MENVPRO

Co-funded by the
Erasmus+ Programme
of the European Union

Environmental Statistics

Dr. Silvia Portarena

National Research Council (CNR)
Research Institute on Terrestrial Ecosystems (IRET)
Porano (TR), Italy

The Environmental
Science Education
for Sustainable Human Health

7 September 2021

UNIVERSITÀ DEGLI STUDI DELLA TUSCIA
GIRAF PM
NAS RA INTERNATIONAL SCIENTIFIC EDUCATIONAL CENTER
MARTIN-LUTHER-UNIVERSITÄT HALLE-WITTENBERG
UNIVERSITY OF GEORGIA
TÉCNICO LISBOA
GRENA GEORGIAN RESEARCH AND EDUCATIONAL NETWORKING ASSOCIATION

Co-funded by the
Erasmus+ Programme
of the European Union

Environmental Statistics

Environmental data

Natural observations (climate models) and experimental measurements (e.g. pollution analyses, ecological data)

Statistics

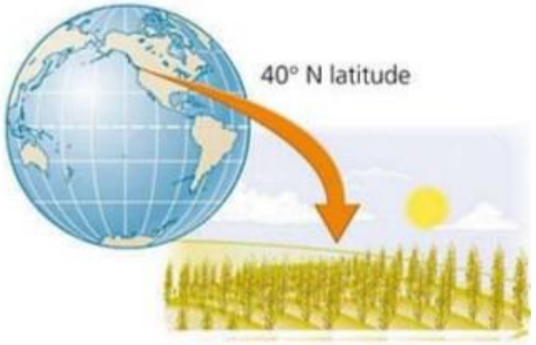
Statistics is a way to get information from data - advanced data analysis

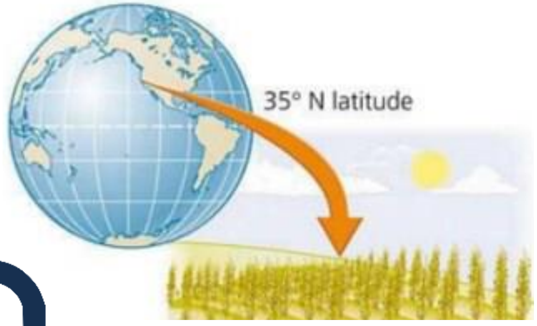
Developing solutions to environmental problems

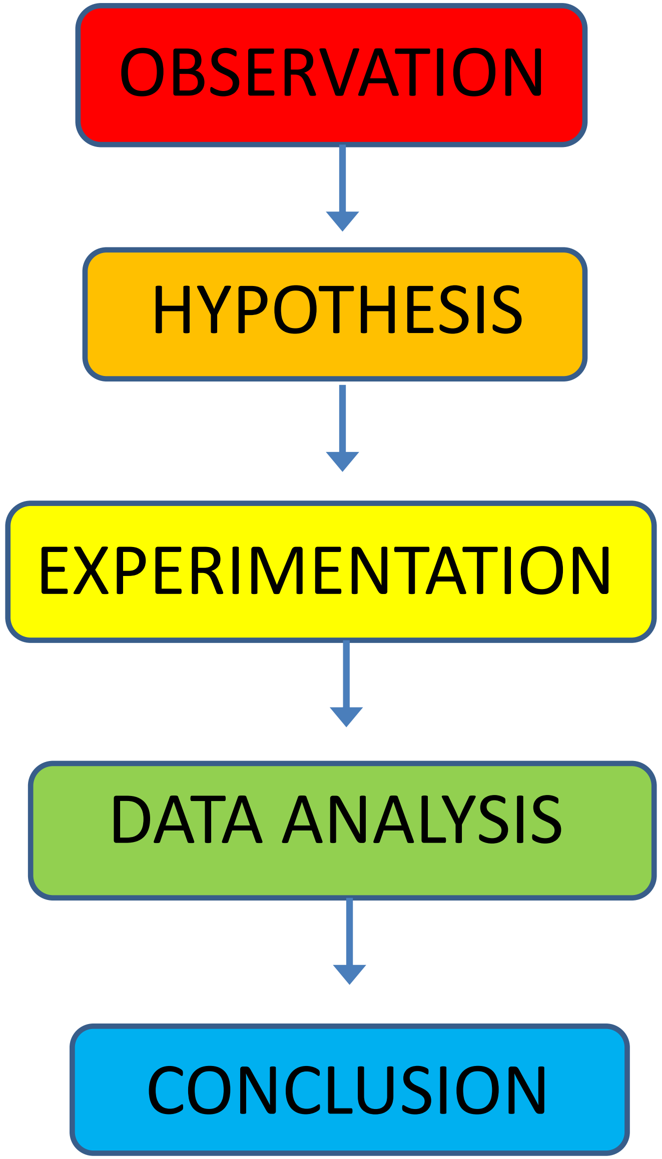
- How does the natural world work?
- How does the natural environment affect us?
- How do we affect our environment?

ENV PRO

Statistics and scientific method







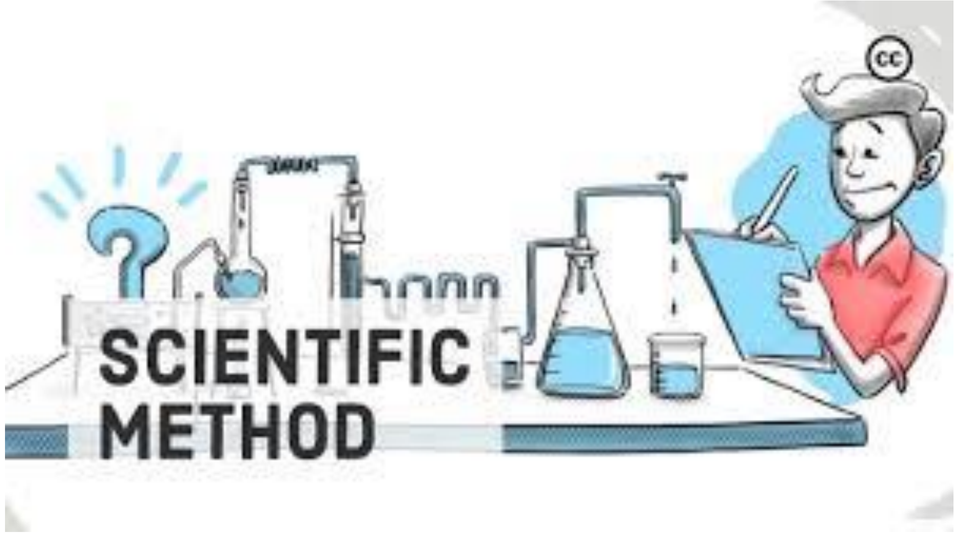
OBSERVATION

HYPOTHESIS

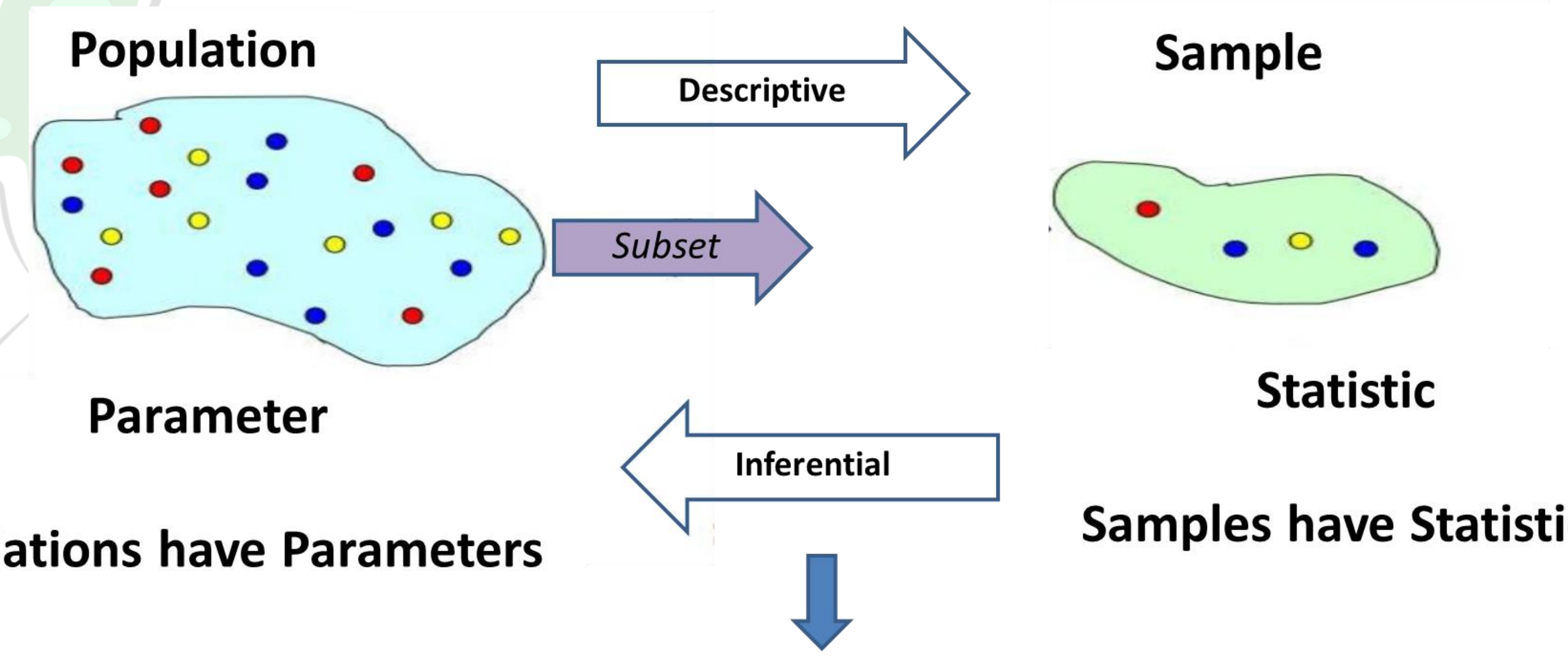
EXPERIMENTATION

DATA ANALYSIS

CONCLUSION



Operative approaches



Population

Descriptive

Subset

Sample

Parameter

Statistic

Inferential

Populations have Parameters

Samples have Statistics

Is the hypothesis true?

Descriptive Statistics

Summarize/extract information from many numbers to lesser number of parameters

Types of descriptive statistics:

- **Organize Data**
  - Tables
  - Graphs
- **Summarize Data**
  - Central Tendency analyses
  - Variation analyses

EXPLORATIVE

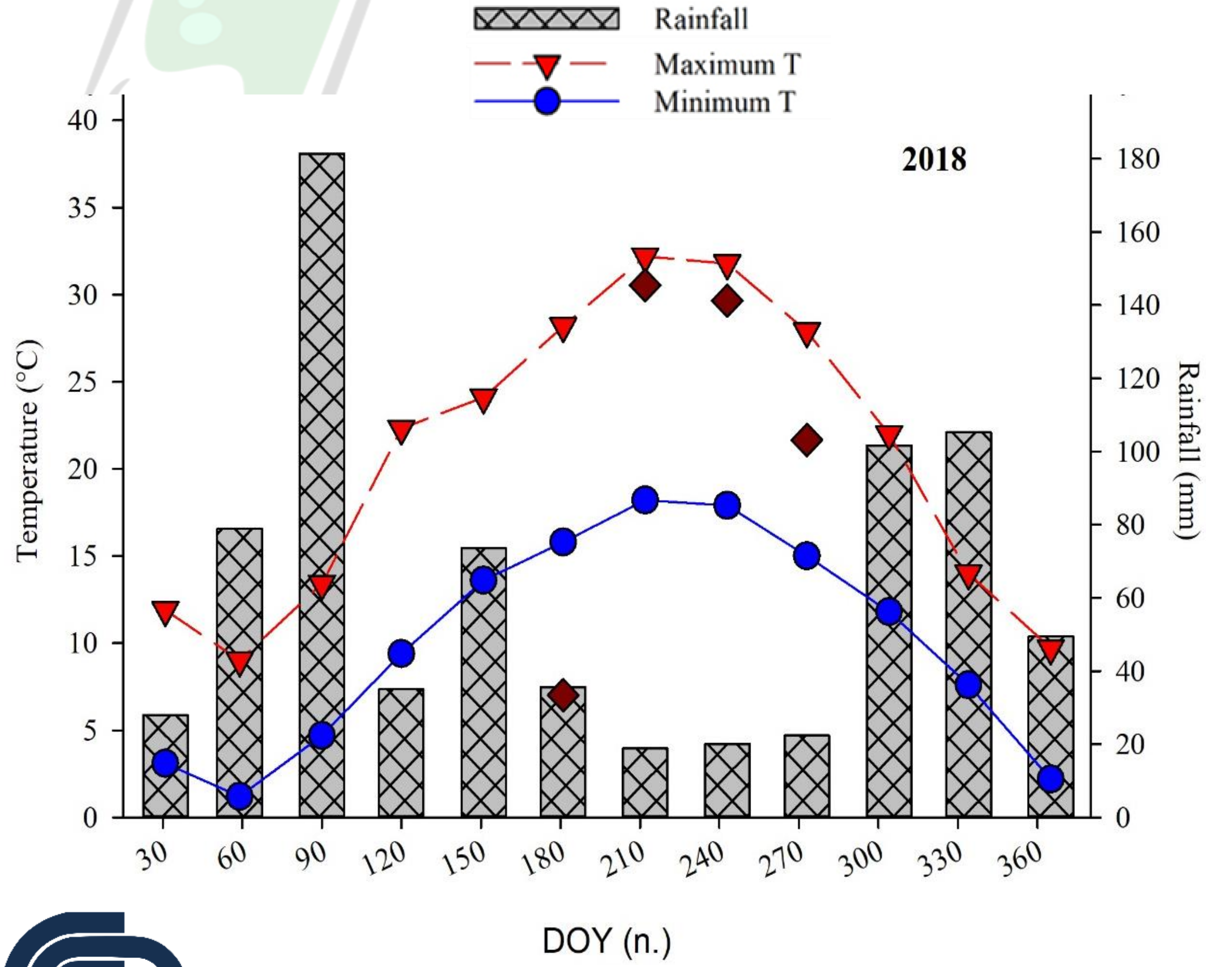
ENV PRO

Descriptive Statistics: Organize Data

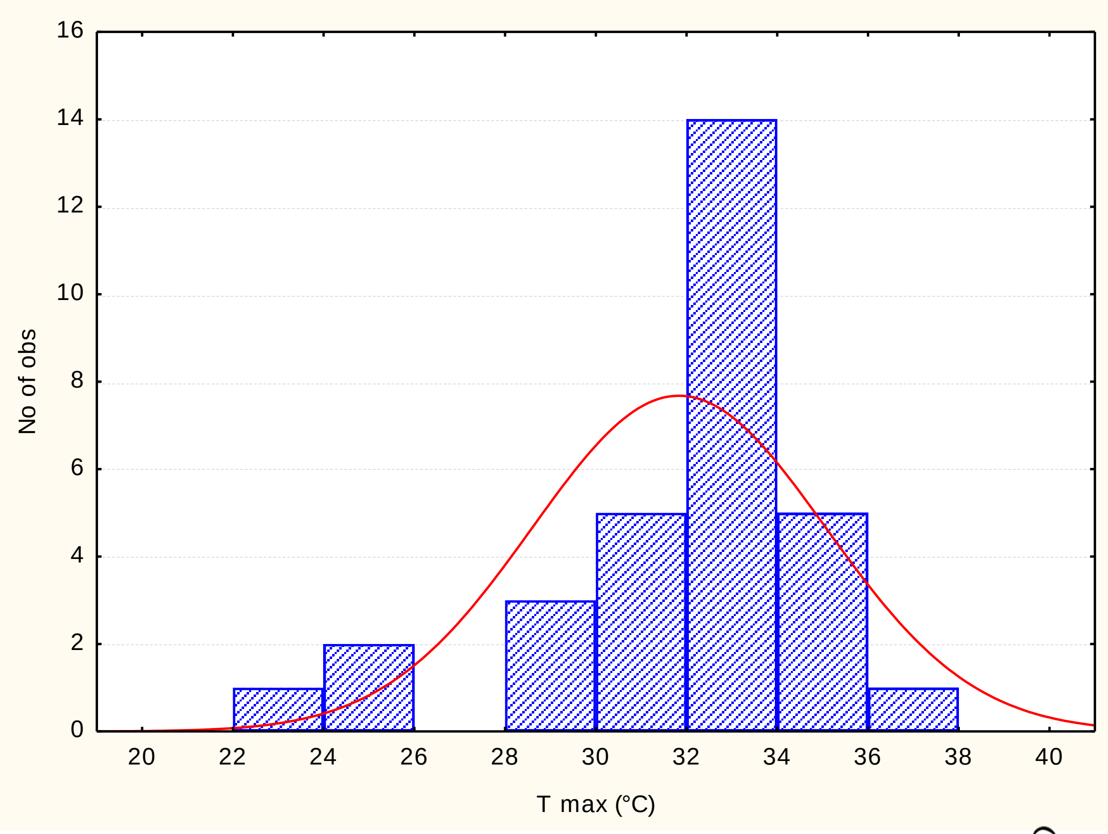
- Tables
  - Frequency Distributions
  - Relative Frequency Distributions
- Graphs
  - Bar Charts
  - Histograms
  - Scatter Plots
  - Box-whisker plots

Graphs…

Bar Chart

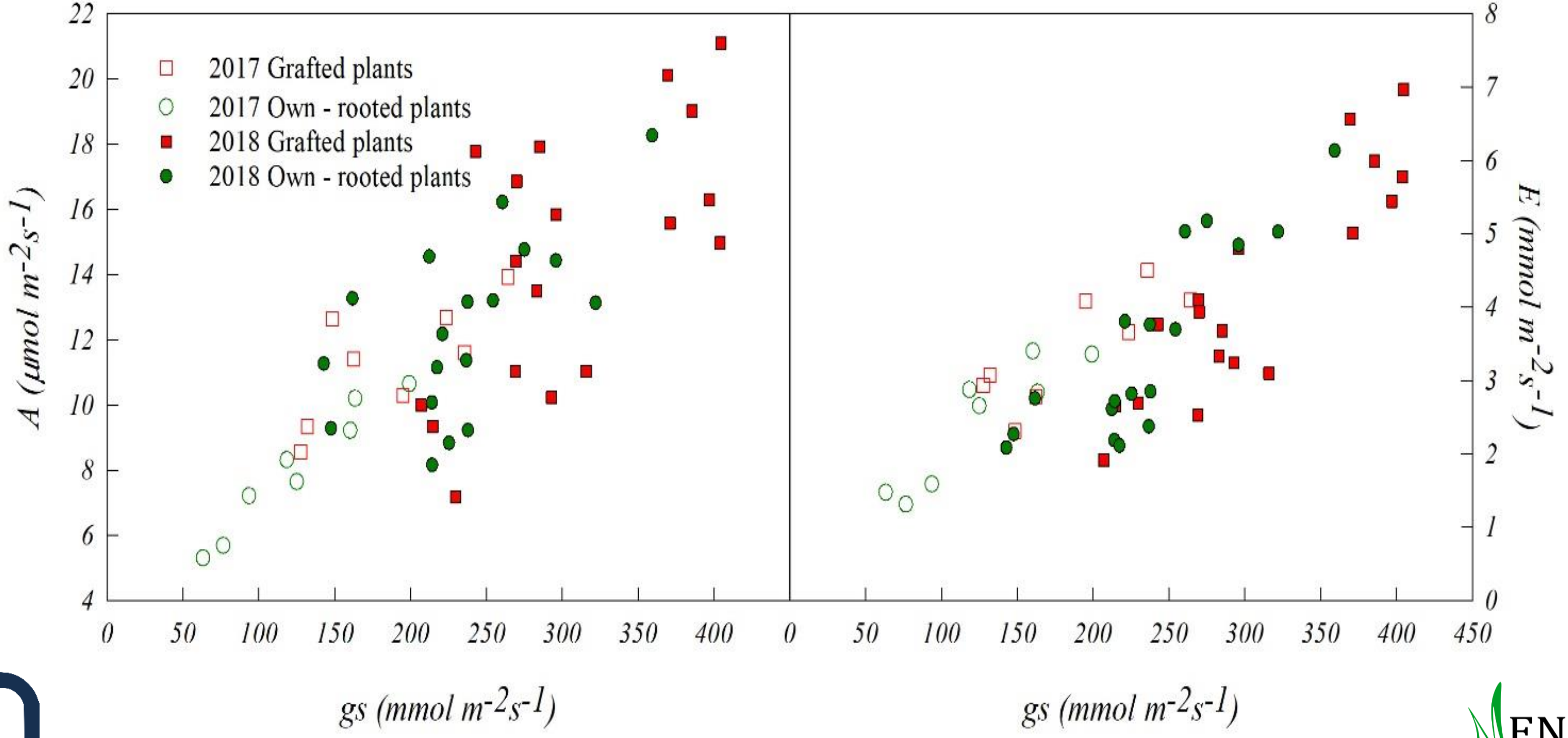


Histogram



Graphs…

Scatterplots



Descriptive Statistics: summarizing data

- **Central Tendency** (or Groups’ “Middle Values”)
  - Mean
  - Median
  - Mode

- **Variation** (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Summarizing Data – Central tendency

Mean

Most commonly called the "average."

Add up the values for each case and divide by the total number of cases.

n = total number of cases

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Summarizing Data – Central tendency

Median

The median is the middle observation that divides a distribution into two equal halves (The 50th percentile).

Example - number of counts in the main peaks of two spectra

Spectrum A	Spectrum B	
8000	7000	
9000	7500	
11000	8000	
12000	8500	***median***
15000	11000	
18000	18000	
20000	39000	
13286	14143	***mean***

Summarizing Data – Central tendency

Mode

The most frequently occurring observation

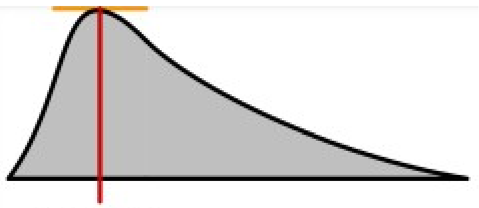
- In a set of a discrete data:

80 87 89 93 93 96 97 98 102 103 105 106 **109 109 109** 110 111 115 119 120 127 128 131 131 140 162

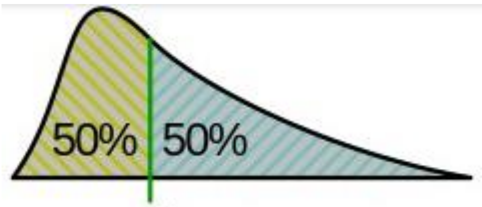
The Mode = 109

It is possible to have more than one mode!

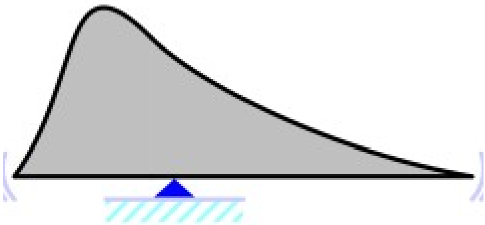
- In the case of continuum variables: the value associated with the maximum probability



mode



median



mean

Summarizing Data – Variation

Range

It is the difference between a minimun and a maximum value

Data set A

102
128
131
98
140
93
110

115
109
89
106
119
97

Range = 140 - 89 = 51

Data set B

127
131
96
80
93
120
109

162
103
111
109
87
105

Range = 162 - 80 = 82

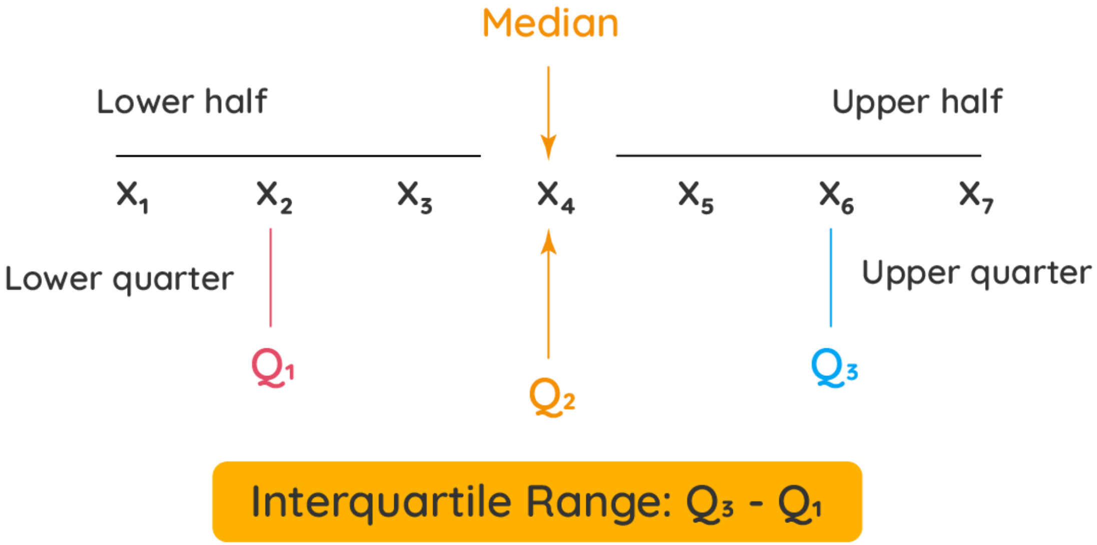
Summarizing Data – Variation

Interquartile Range

The interquartile range is the range between the 75th percentile and the 25th percentile.

A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

25th percentile (Q_1) is a quartile that divides the first 1/4 of cases from the latter 3/4.
75th percentile (Q_3) is a quartile that divides the first 3/4 of cases from the latter 1/4 .



Interquartile Range: $Q_3 - Q_1$

Summarizing Data – Variation

Variance

The sum of square deviations from mean divided by n-1

A measure of the spread of the recorded values on a variable.
The larger the variance, the further the individual cases are from the mean.

The smaller the variance, the closer the individual scores are to the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- n = number of samples
- n-1 = degrees of freedom
- $\bar{x}$ = average of samples

Summarizing Data – Variation

Standard deviation

It is the square root of the variance. It is in the same unit as data

Data set of 20 measurements

1	8
2	10
3	14
4	12
5	8
6	9
7	12
8	6
9	11
10	9

11	8
12	6
13	13
14	7
15	5
16	14
17	8
18	9
19	6
20	8

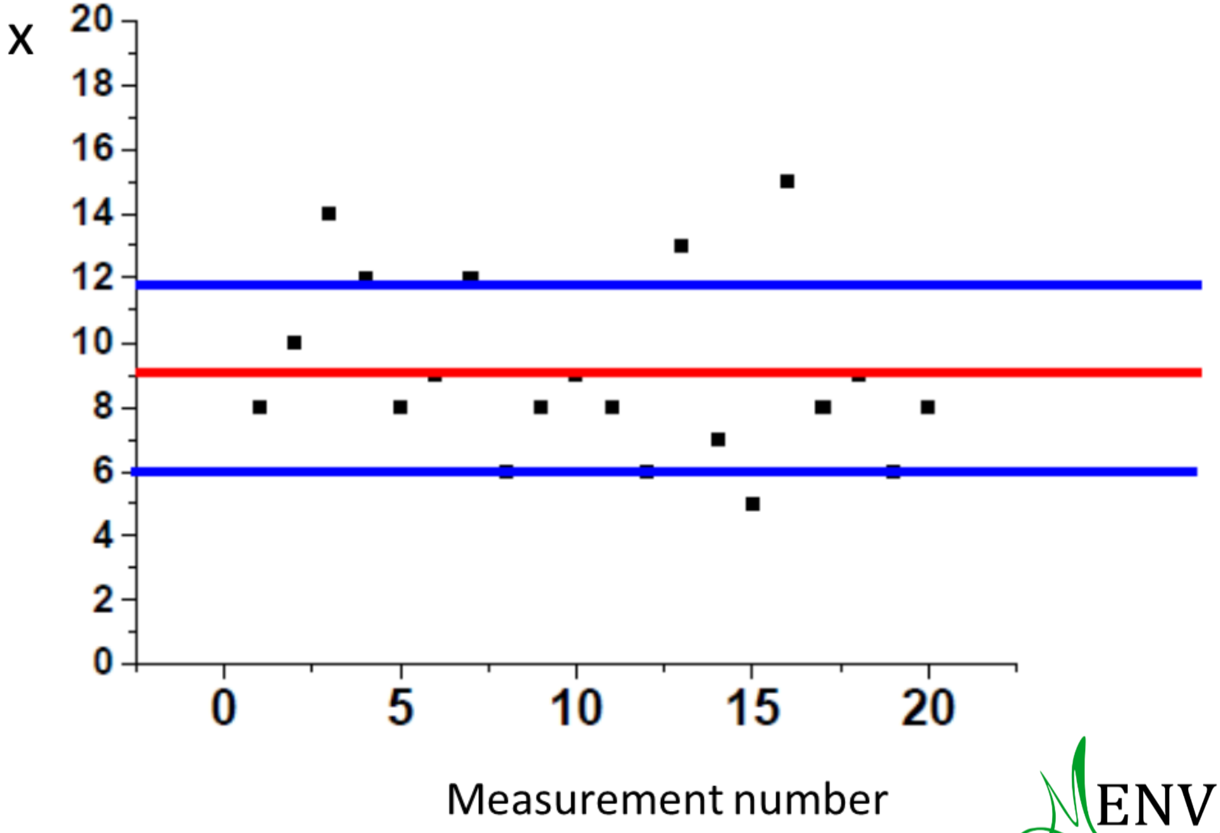
n = 20

$\bar{x}$ = 9.15

σ = 2.72

X = 9 ± 3

X = 9.2 ± 2.7

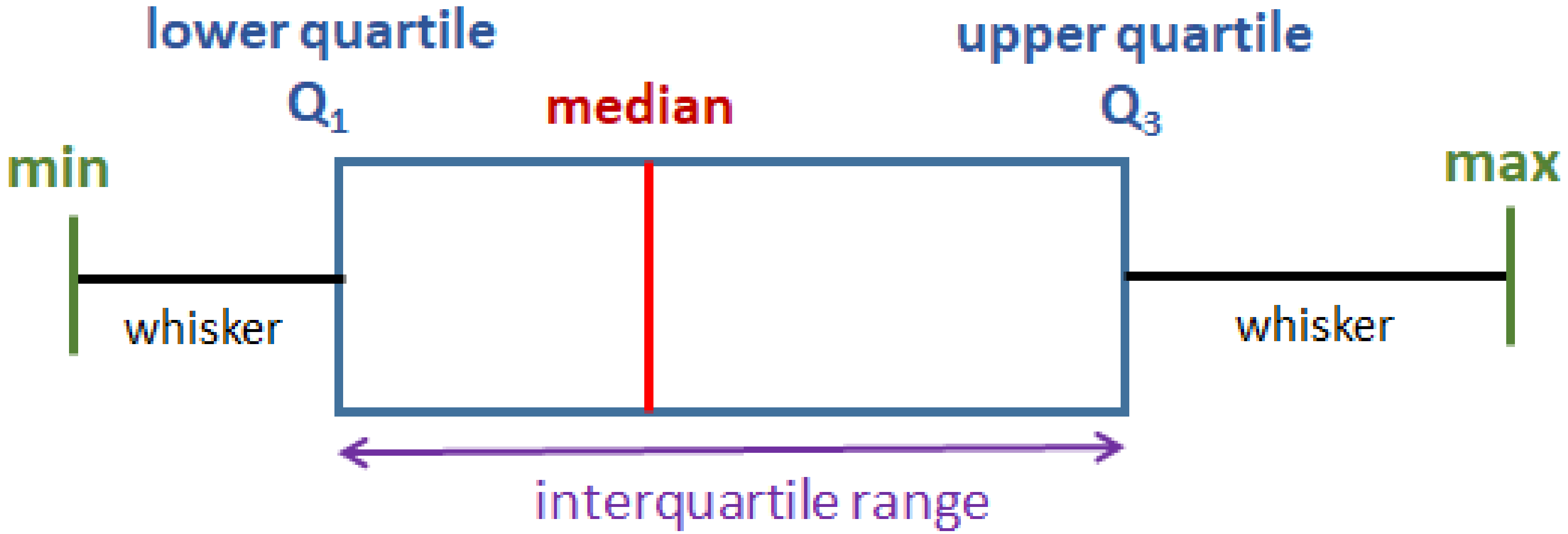


Graphic summarizations: Box-whisker plot

It is a way of displaying the five-number summary of a data set:

- It shows:

minimum and maximum values, lower (Q_1) and upper quartiles (Q_3), median



It is useful to campare the distributions of different variables

Inferential Statistics

The process of using data analysis to infer properties of a population

- It is assumed that the observed data set is sampled from a larger population.

Steps:

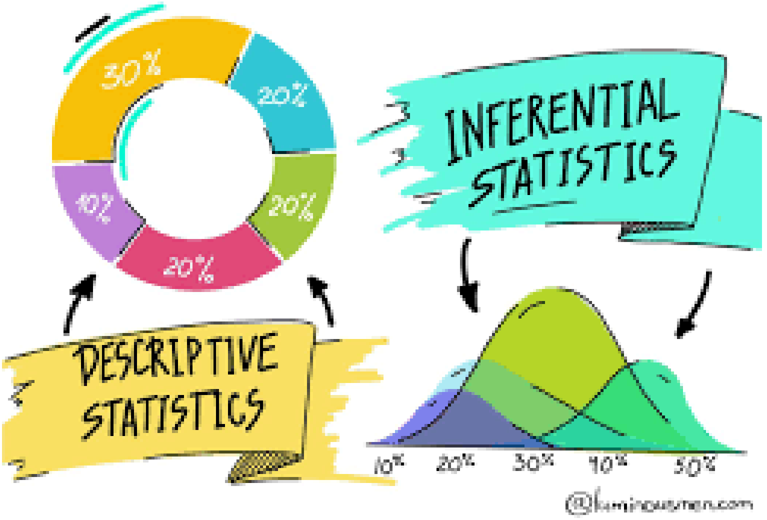
- Select a statistical model of the process that generates the data
- Formulate the hypothesis
- Select the confidence level
- Test the hypothesis on the basis of experimental data (the experiments have to be reproducible)
- Reject or fail to reject the hypothesis
- Built new theories or hypothesis

INFERENTIAL STATISTICS IS COMMON IN ENVIRONMENTAL SCIENCE!

Inferential statistics and probability

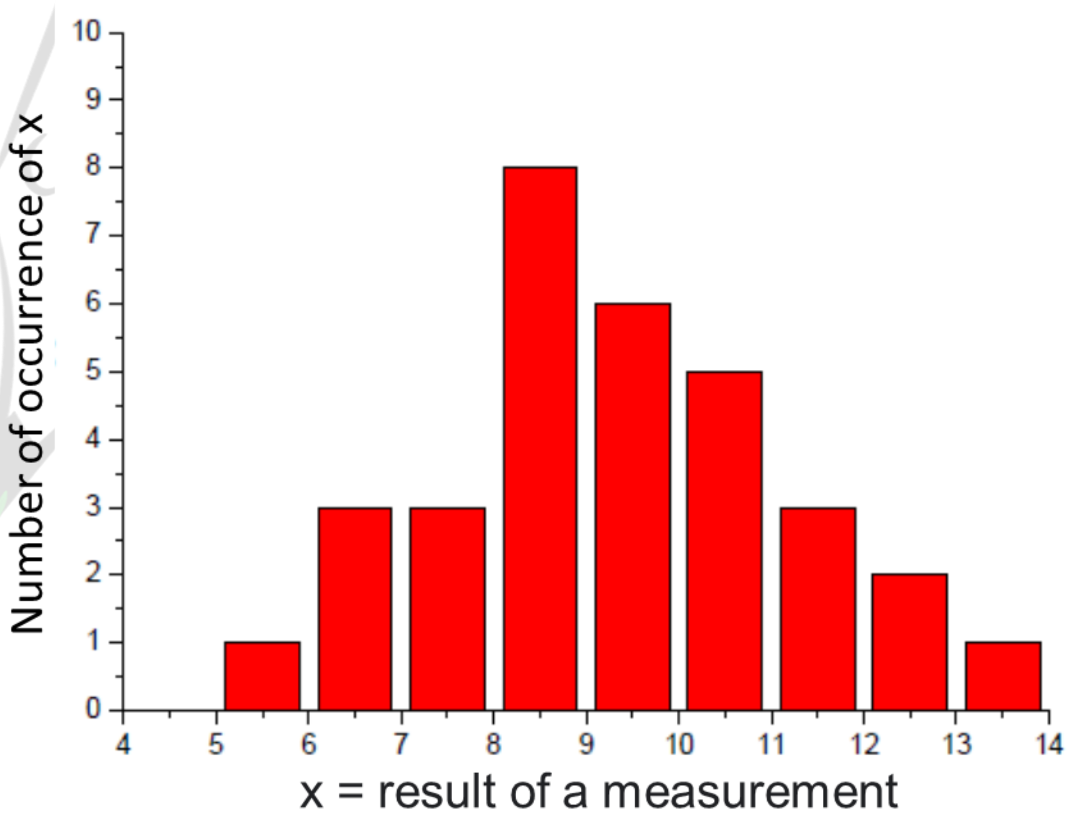
Inferential statistics is built on the foundation of probability theory

- Used to quantify likelihood that particular values occur
- Used to represent risk or uncertainty in experimental applications
- Used to estimate the validity of a hypothesis



The importance of probability

A sample needs to be representative of the population

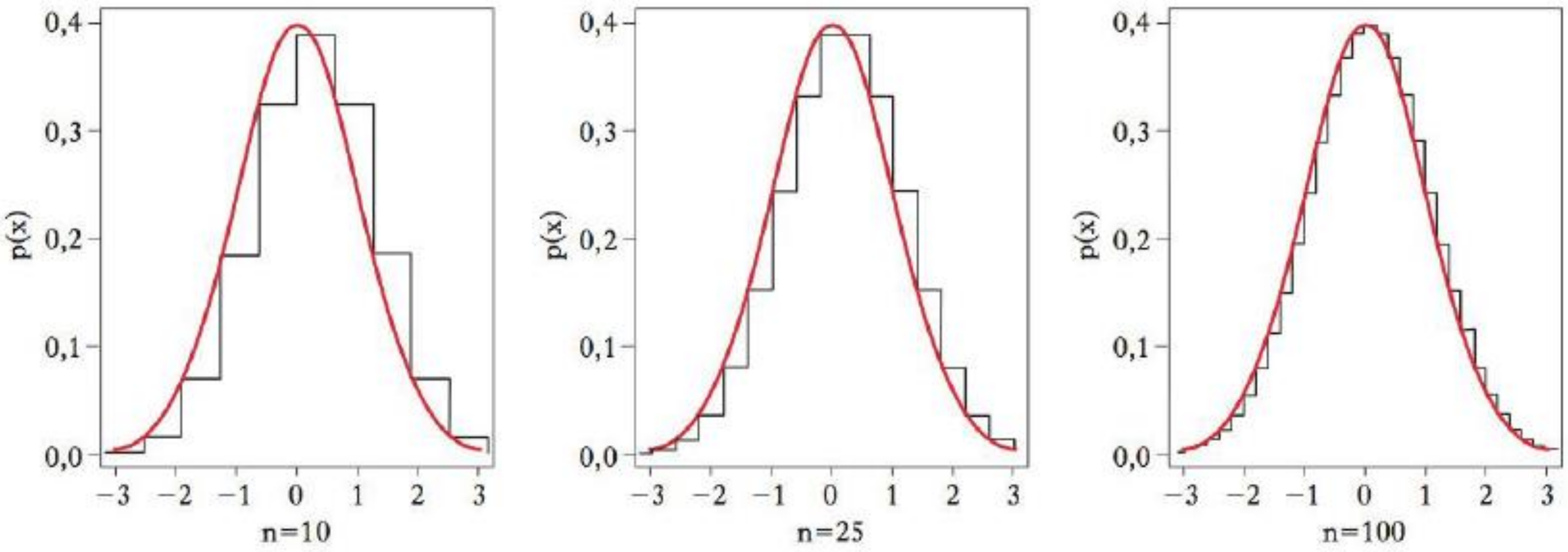


Frequency distribution.

- Since errors always exist in our experimental values we can only deal with probability of being correct or incorrect in our results.
- A histogram may suggest a theoretical distribution of a measured variable
- The height of each bar gives the probability that variable x lies in the given range

Distribution limit

By increasing the number of measures in many cases the histogram begin to take a well-defined form that does not change over time



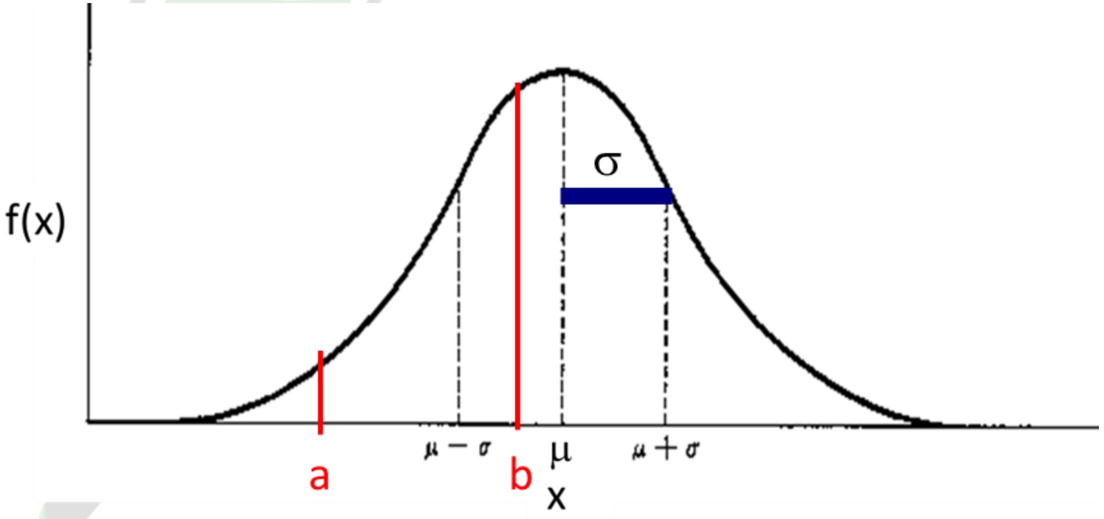
The Normal Distribution is a density curve based on the following formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- It's completely defined by two parameters: mean (μ, location); and standard deviation (σ, spread)
- The normal distribution is symmetrical.
- The mean, median, and mode are all the same.

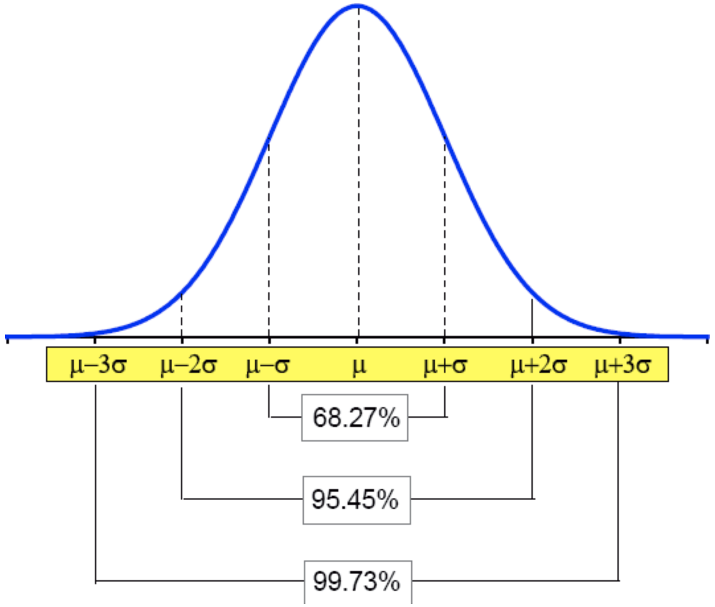
Normal or Gaussian distribution

The most popular distribution for continuous random variables



The probability of observing a random variable x in the range [a,b] is the integral between a and b of its normal distribution

$$P(a < x < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



Confidence level

Gives the probability with which an estimated interval will contain the true value of the parameter.

- σ ≤ μ ≤ σ	P(σ) =68%
- 2σ≤ μ ≤ 2σ	P(2σ) =95%
- 3σ ≤ μ ≤3σ	P(3σ) =99.7%

Checking of normality

Most of the statistical analyses make assumptions about normality of data

There are two main methods of assessing normality: graphical and numerical (including statistical tests)

Graphs

- Histogramm
- Box plot
- Normal probability (Q-Q) plot

A Q-Q plot is a scatterplot created by plotting the quantiles (percentiles) from experimental data set against that of a normal distribution

Statistical formal tests

- The Kolmogorov–Smirnov test (n> 50 samples)
- The Shapiro–Wilk test (n< 50 sample)

(are most widely used methods to test the normality of the data)

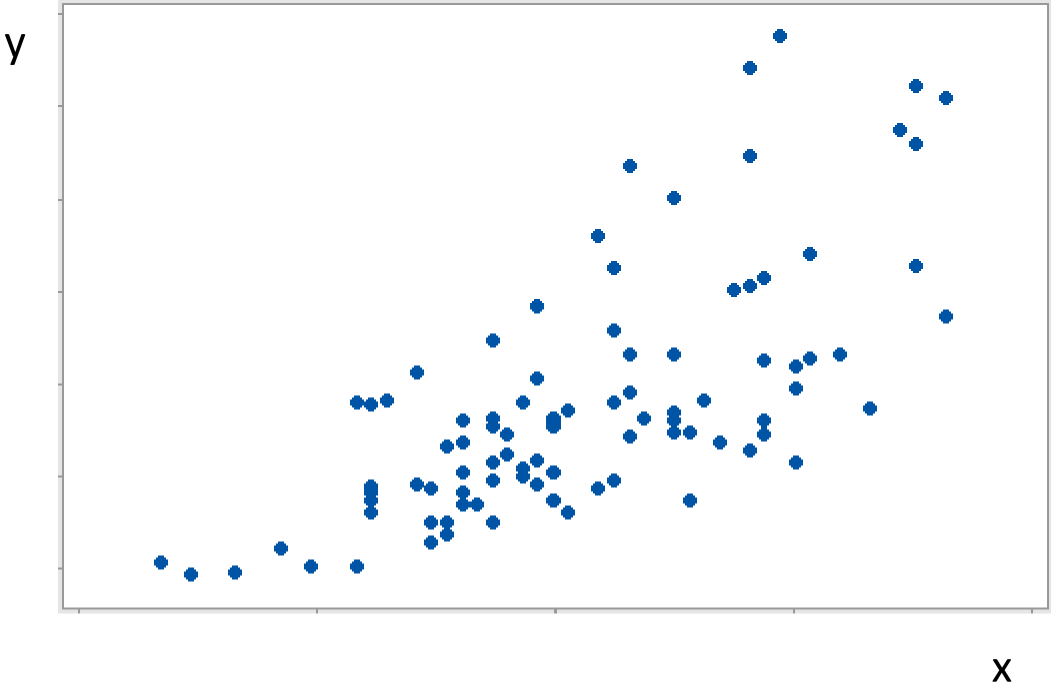
For both of the above tests, null hypothesis states that data are taken from normal distributed population. When P > 0.05, null hypothesis accepted and data are called as normally distributed.

Statistical tests have the advantage of making an objective judgment of normality but have the disadvantage of sometimes not being sensitive enough at low sample sizes

Measures of Relationship

Topics Covered:

- Is there a relationship between *x* and *y*?
- What is the strength of this relationship
- Can we describe this relationship and use this to predict *y* from *x*?
- Is the relationship we have described statistically significant?



↓

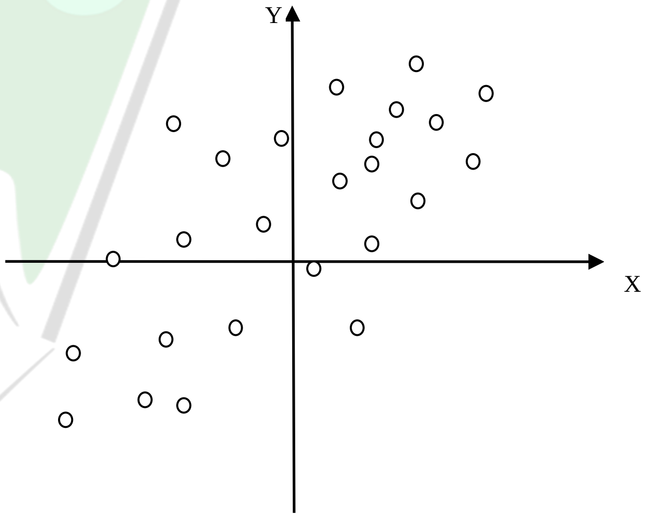
Correlation and Regression analyses

The relationship between x and y

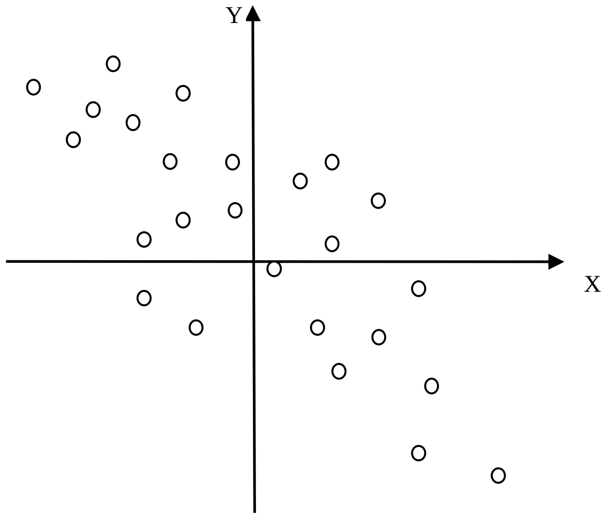
- Correlation: is there a relationship between 2 variables?
- Regression: how well a certain independent variable predict a dependent variable?

- CORRELATION ≠ CAUSATION
  - In order to infer causality: manipulate the independent variable and observe the effect on the dependent variable

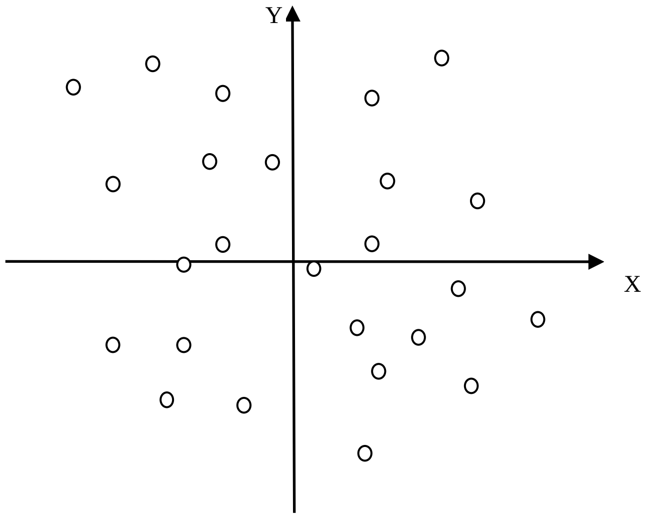
Scatterplots



Positive correlation



Negative correlation



No correlation

The correlation analysis gives the strength and direction of the relationship between variables

Measures of Correlation

Goal: determine the degree of relationship between variables

1) Covariance

2) Pearson Correlation Coefficient (r)

Measures of Correlation

1) Covariance

It is a measure of the joint variability of two random variables

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n}$$

Problem: the value obtained by covariance is dependent on the size of the data n

- When X $\uparrow$ and Y $\uparrow$: cov (x,y) = pos.
- When X $\uparrow$ and Y $\downarrow$: cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

Measures of Correlation

2) Pearson correlation coefficient (r)

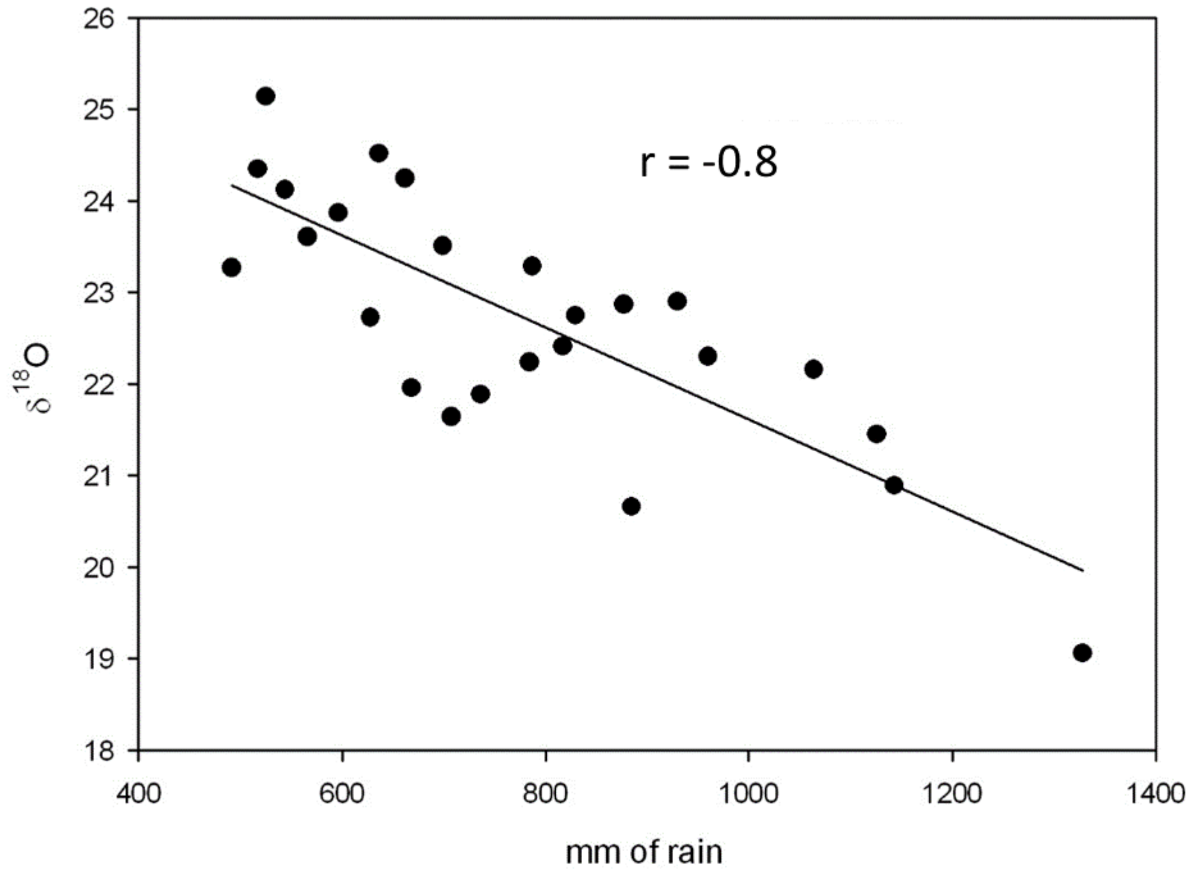
$$r = \frac{\sum(Xi-\bar{X})(Yi-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}} = \frac{\operatorname{cov}(x,y)}{\sigma_x\,\sigma_y}$$

(σ = standard dev of sample)

- Pearson’s r: standardizes the covariance value.
- Divides the covariance by the multiplied standard deviations of X and Y
- r , dimensionless number (covariance standardized by variances of single variables)
- **r takes values fom -1 (perfect negative correlation) to 1 (perfect positive correlation). r=0 means no correlation**

Example

Negative correlation between δ^{18}O values of olive oil samples and the amount of precipitation of native areas of the olive plants.



Decreasing isotope compositions of olive oil with increasing precipitations

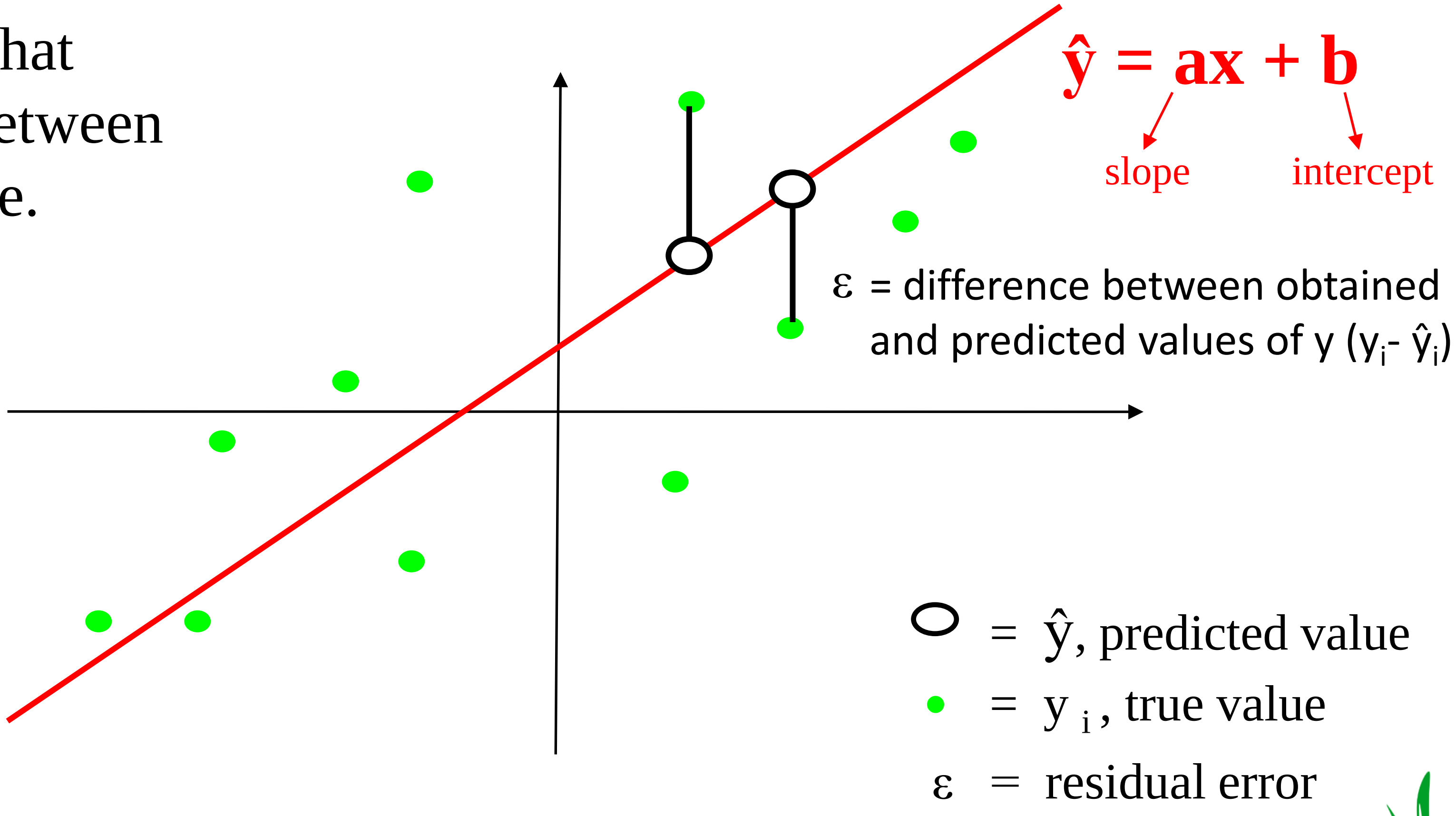
Regression

- Correlation tells you if there is an association between x and y but it doesn’t describe the relationship or allow you to predict one variable from the other.

- To do this we need REGRESSION!

Best-fit line for the data

- Aim of linear regression is to fit a straight line, $\hat{y} = ax + b$, to data that gives best prediction of y for any value of x
- This will be the line that minimises distance between data and fitted line, i.e. the residuals



ε = difference between obtained and predicted values of y (y_i- $\hat{y}_i$)

○ = ŷ, predicted value

● = y_i, true value

ε = residual error

Least Squares Regression

- To find the best line we must minimise the sum of the squares of the residuals (the vertical distances from the data points to our line)

Model line: $\hat{y} = ax + b$ a = slope, b = intercept

Residual (ε) = $y - \hat{y}$

Sum of squares of residuals = $\Sigma\,(y_i - \hat{y}_i)^2$

- we must find values of a and b that minimise

$$\Sigma\,(y_i - \hat{y}_i)^2$$

Minimising sums of squares

- Minimise $\Sigma(y_i- \hat{y}_i)^2$, which is $\Sigma(y_i-ax_i+b)^2$

- Minimum $\Sigma(y_i-ax_i+b)^2$ is at the bottom of the curve where the gradient is zero – and this can found with calculus

- Take partial derivatives of $\Sigma(y_i-ax_i-b)^2$ respect to parameters *a* and *b*, set them equal to zero; and solve, giving:

$$\boxed{a = \frac{\sum_{i=1}^{n}(X_i - \bar{x})}{\sum_{i=1}^{n}(X_i - \bar{x})(yi - \bar{y})} = \frac{Cov\ (x,y)}{Var\ (x)}}$$

Slope

$$\boxed{b = \bar{y} - a\ \bar{x}}$$

intercept

Regression line always goes through the mean point: $\bar{x}$ and $\bar{y}$

Model evaluation -Significance testing

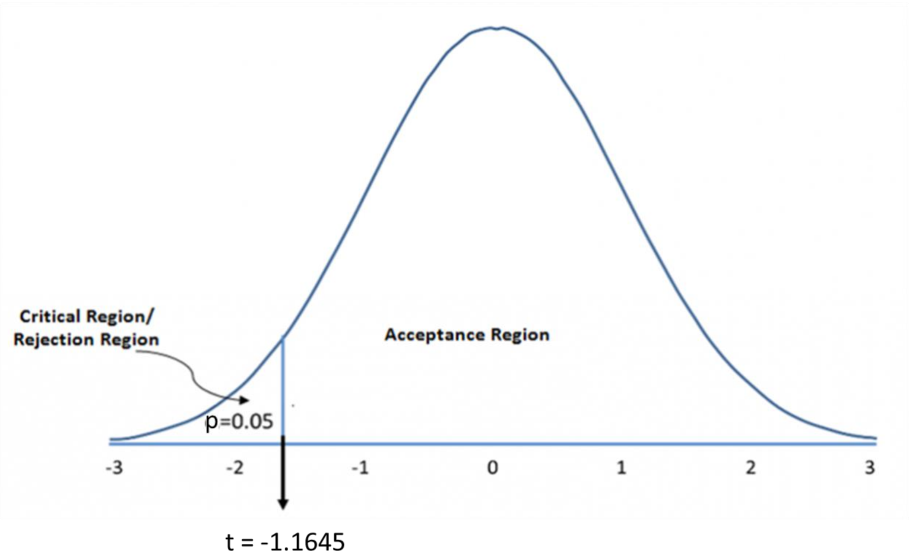
Test of null hypothesis

H_0: a = 0 (no linear relationship)

H_1: a $\neq$ 0 (linear relationship does exist)

- Statistical t test:

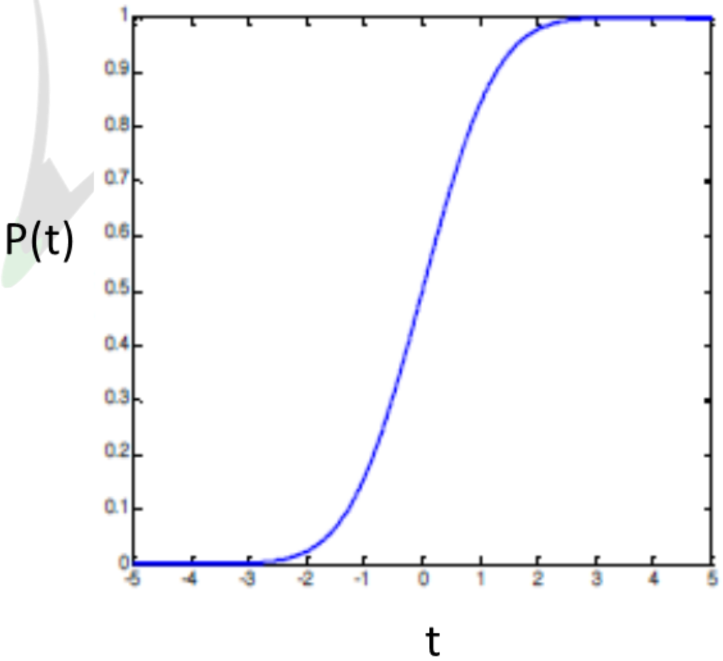
$$t_{n-2} = \frac{a-0}{\sigma_a} \longrightarrow f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(t)^2}{2}} dt$$



Model evaluation -Significance testing

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(t)^2}{2}} dt$$

$$P(t\ ) = \int_{-\infty}^{t*} \frac{1}{\sqrt{2\pi}} e^{-\frac{(t)^2}{2}} dt$$



I reject H_0 if the probability value is below 0.05 (p- value)

The lower the p-value, the stronger the evidence that the null hypothesis is false

↓

Linear relationship does exist

An appropriate statistical software can compute the analysis, giving the right statistical parameters

Data transformation to linearity

- I plot the data and note that they follow a non-linear pattern

- Many non-linear curves can be put into a linear form by appropriate transformations (mathematical functions) of the either:
  - the dependent variable Y or
  - the independent variable X
  - or **both**.

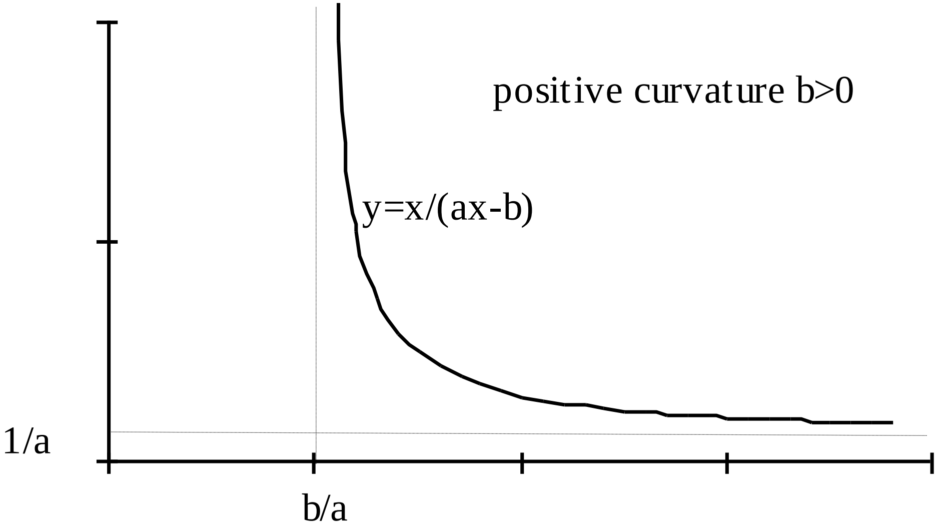
Intrinsically Linear (Linearizable) Curves

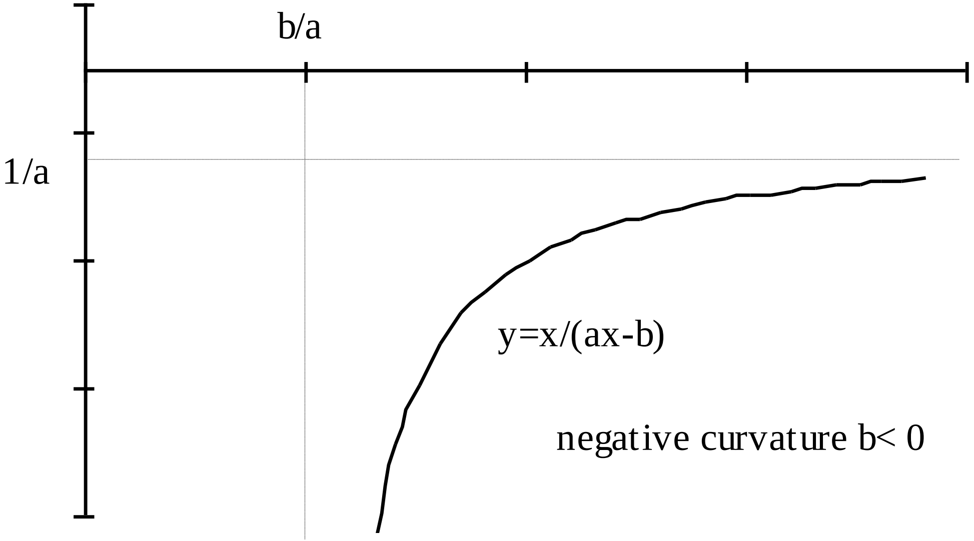
1. Hyperbolas

$y = x/(ax-b)$

Linear form: $1/y = a - b(1/x)$ or $\boldsymbol{Y = \beta_0 + \beta_1 X}$

Transformations: $Y = 1/y$, $X = 1/x$, $\beta_0 = a$, $\beta_1 = -b$





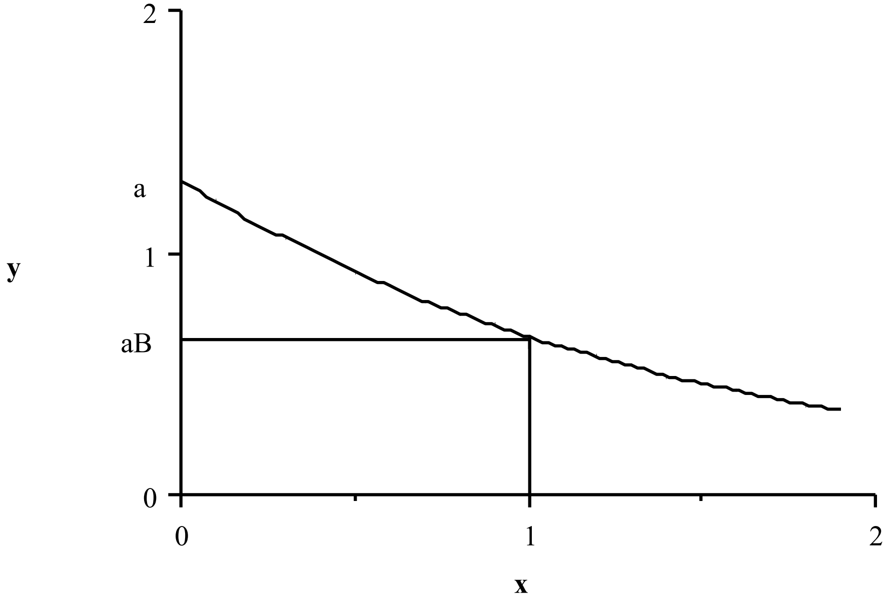
2. Exponential

$y = \alpha\, e^{\beta x} = \alpha B^{x}$

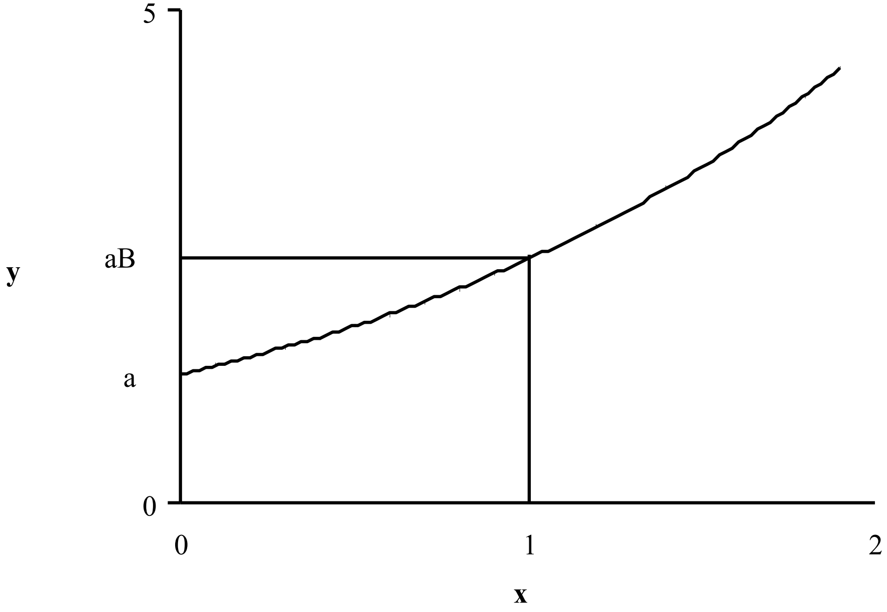
Linear form: $\mathit{ln}\, y = \mathit{ln}\alpha + \beta\, x = \mathit{ln}\alpha + \mathit{ln}B\, x$ or $\mathbf{Y = \beta_0 + \beta_1 X}$

Transformations: $Y = \mathit{ln}\, y,\quad X = x,\quad \beta_0 = \mathit{ln}\alpha,\quad \beta_1 = \beta = \mathit{ln}B$

Exponential (B < 1)



Exponential (B > 1)

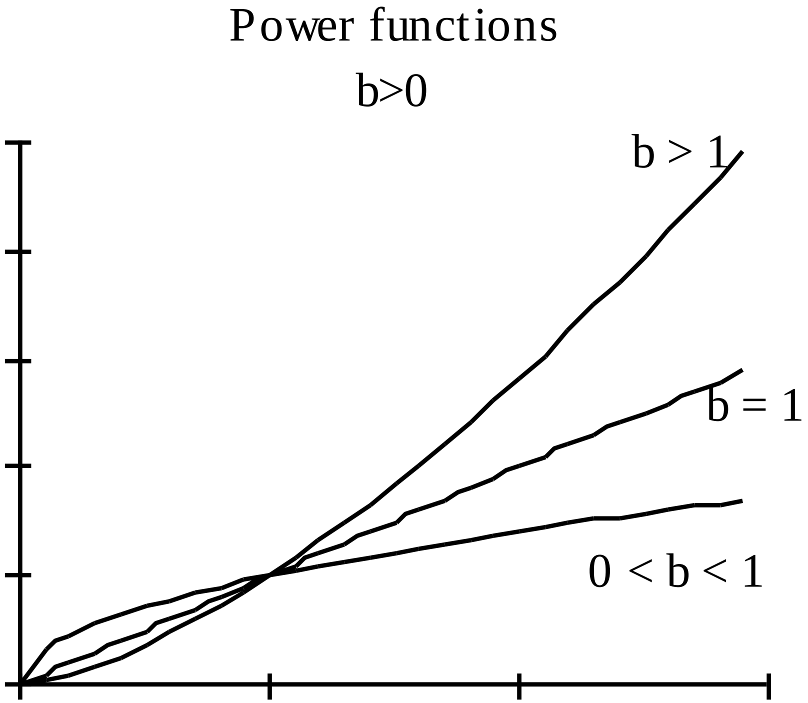


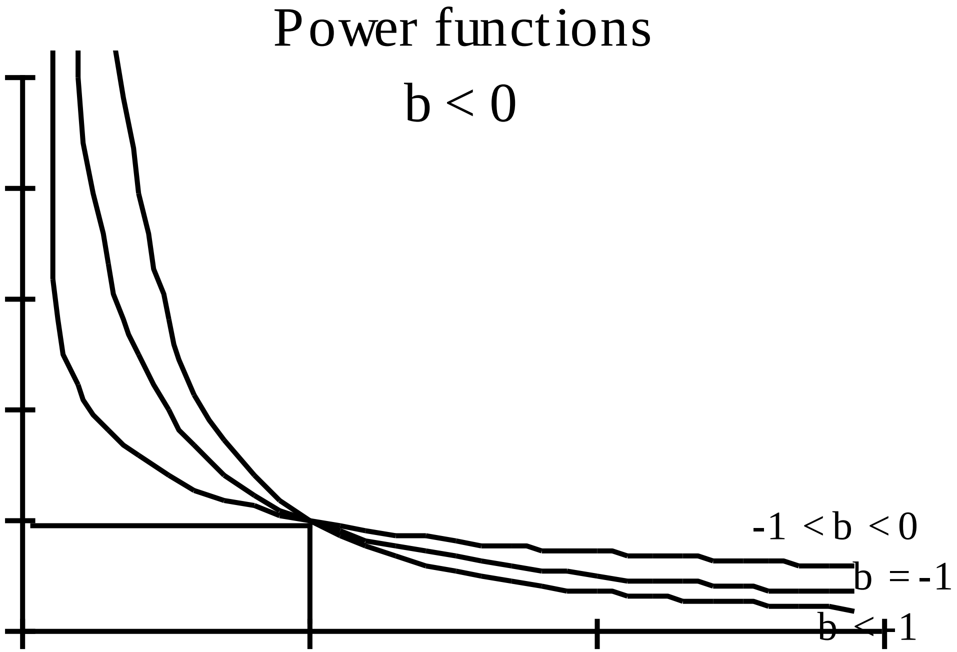
3. Power Functions

$y = a\,x^b$

Linear from: $\ln y = \ln a + b \ln x$ or $\mathbf{Y = \beta_0 + \beta_1 X}$

Transformations: $Y = \ln y$, $X = \ln x$, $\beta_0 = \ln a$, $\beta_1 = b$





More complex…. multivariate methods

Often environmental research involve many variables for each sample (chromatographic analyses, spectroscopic analyses…)

Four main categories:

- **EXPLORATION**

providing insight into structure and relationships among variables

- **CLUSTERING**

grouping of variables that exhibit similar characteristics

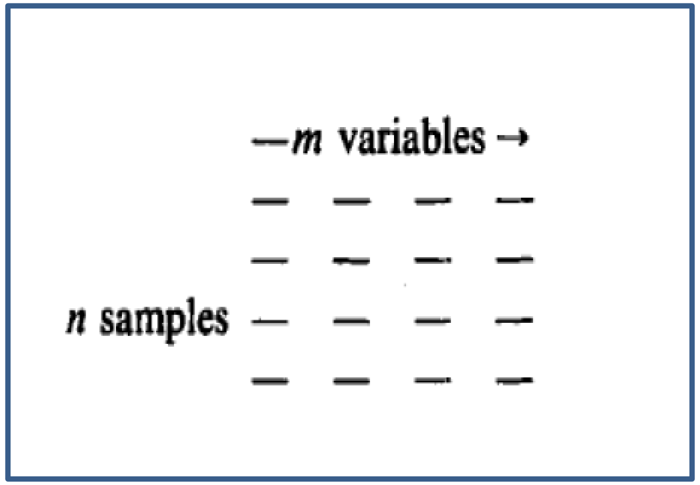
- **CLASSIFICATION**

Classifying variables into predefined groups

- **REGRESSION**

Finding multivariate regressions, where there is more than one response variable

The multivariate data-set is a matrix $M_{n\times m}$



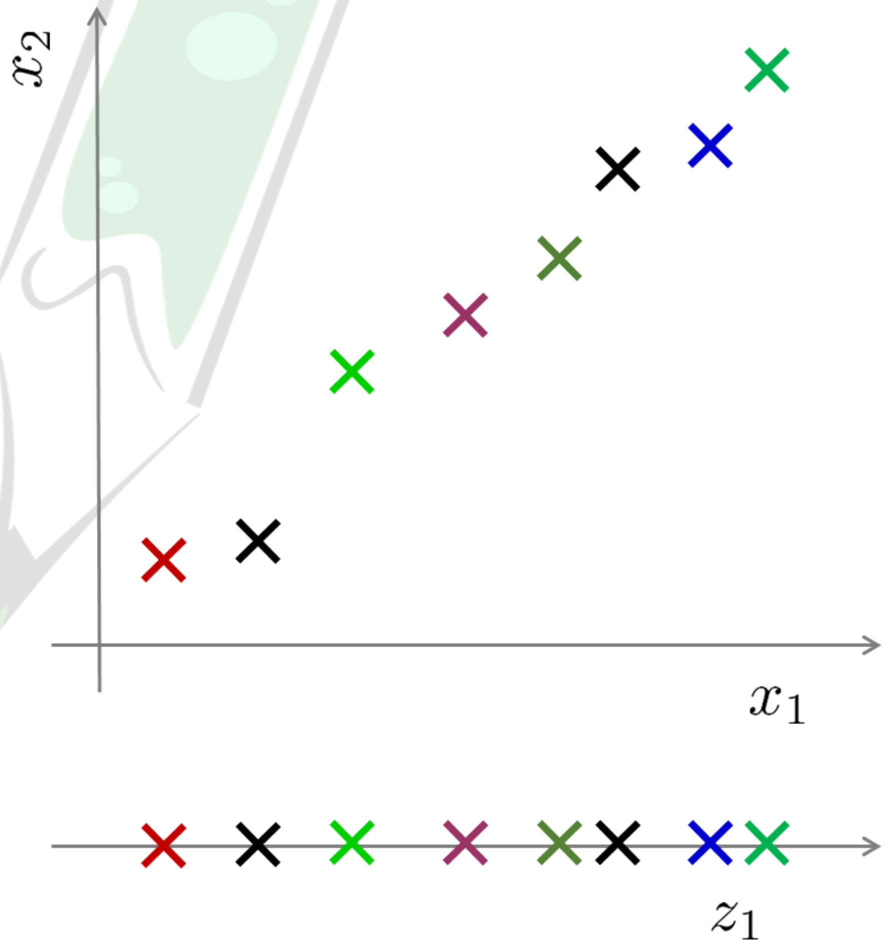
Principal Component Analysis (PCA)

Motivation

- ***Exploration***
  - One way to display patterns in a multivariate data set
- ***Dimensionality reduction***
  - Another way to simplify complex high-dimensional data
  - Summarize data with a lower dimensional vector

- Given data points in *d* dimensions
- Convert them to data points in *r<d* dimensions
- With minimal loss of information

Dimensionality reduction



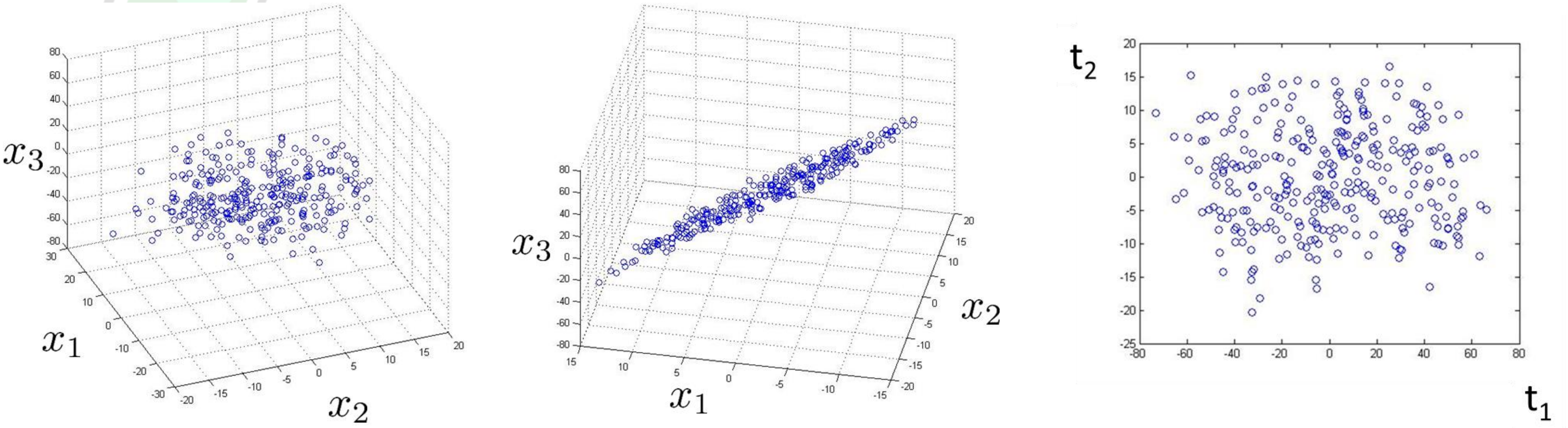
Reduce data from 2D to 1D

$$x^{(1)} \rightarrow z^{(1)}$$
$$x^{(2)} \rightarrow z^{(2)}$$
$$\vdots$$
$$x^{(n)} \rightarrow z^{(n)}$$

For p random variables X_1,…,X_n. the goal of PCA is to construct a new set of N axes in the directions of greatest variability.

Principal Component Analysis (PCA) problem formulation

$3D \to 2D$



Reduce from n-dimension to k-dimension: Find k vectors $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ onto which to project the data, in the directions of greatest variability so as to minimize the projection error.

PCs are linear combinations of the original variables

$$t_{i1} = p_{11}x_{i1} + p_{21}x_{i2}+....+ p_{m1}x_{im} = x_ip_i$$

$$t_{i2} = p_{12}x_{i2} + p_{21}x_{i2}+....+ p_{m2}x_{im} = x_ip_2$$

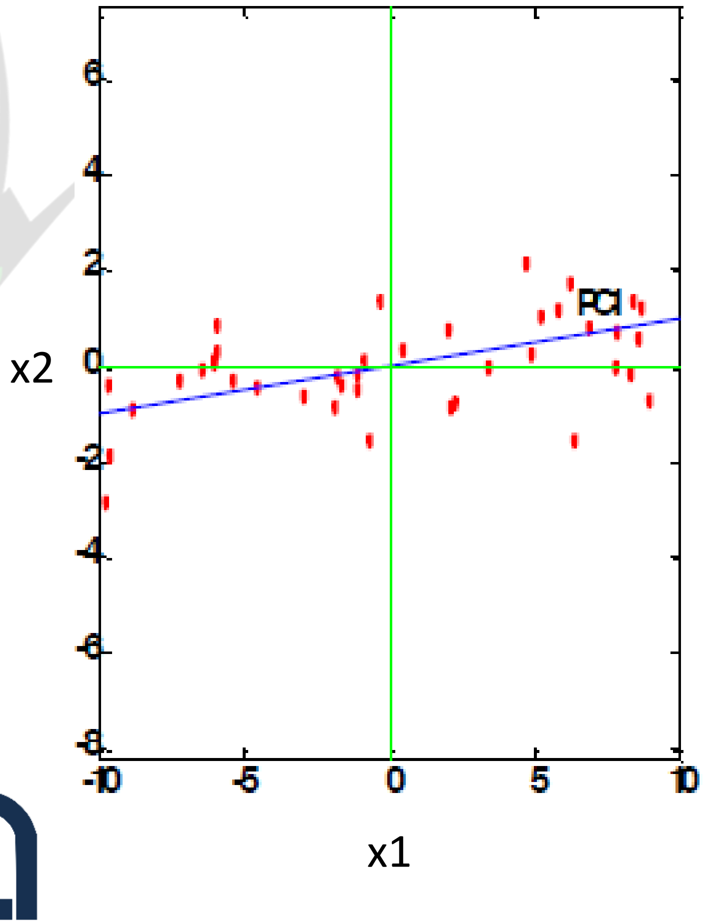
- t_{i1}, t_{i2} are the coordinates of the sample-i on the first and second PC
- x_i the old coordinates corresponding to the measured data
- p_{mi} are the linear combination coefficients

MATRIX FORMULATION: $\underset{n\times f}{\mathbf{T}} = \underset{n\times m}{\mathbf{X}}\ \underset{m\times f}{\mathbf{P}}$

- X is the original matrix of the data set
- T is the score matrix – the sample coordinates in the new PCs
- P is the loading matrix - linear combination coefficients describing the new PCs

PC and variables

- What are the most contributing variables at each PC?



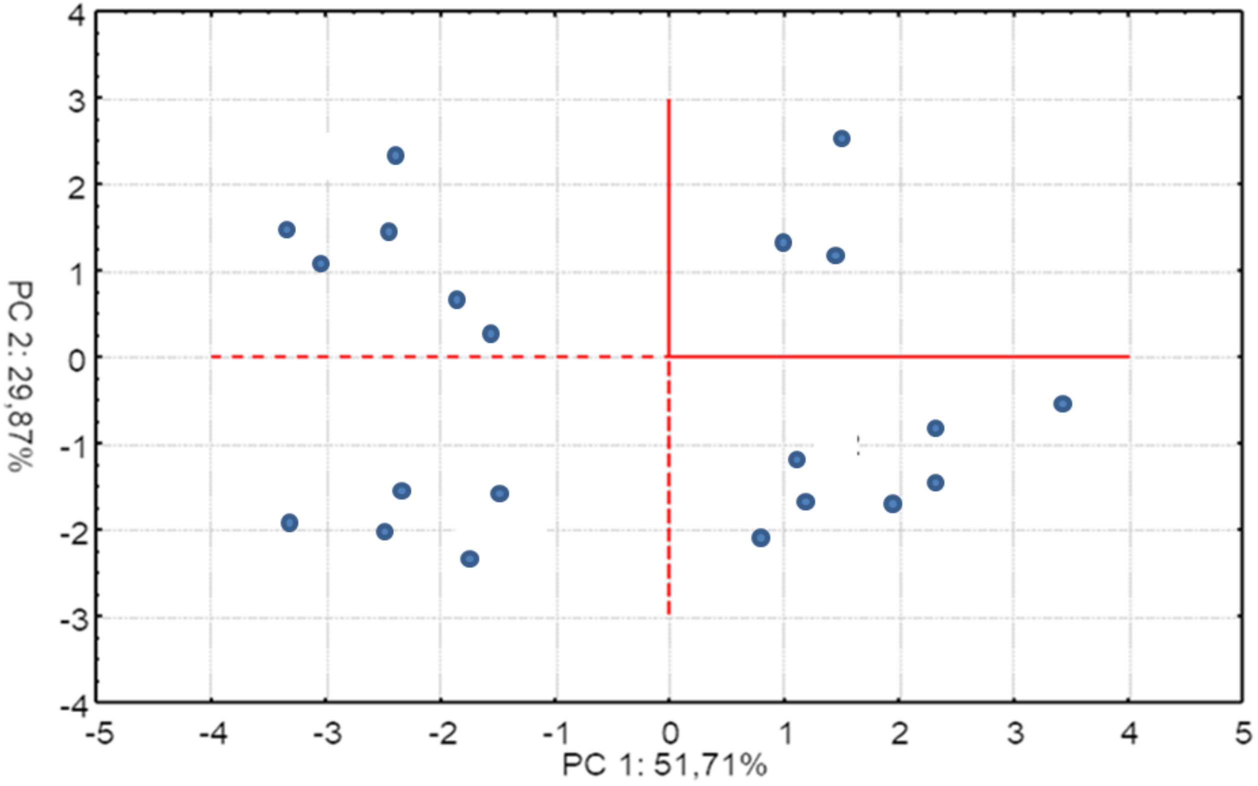
In this graph the PC1 is more similar to the variable x1

In mathematical terms:

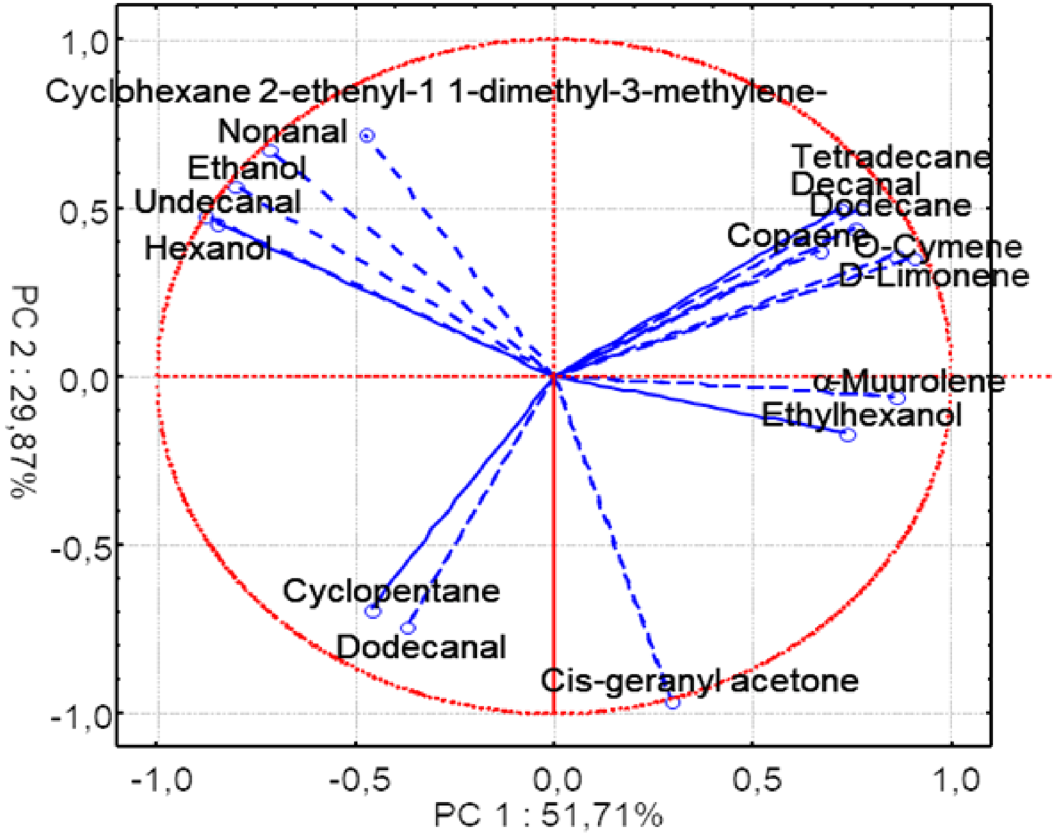
- the **contribution of each variable at each PC** is determined by cos α;
- α: angle between the variable and PC coordinates.
- cos α_i = **loadings** describing the PCs
- -1 < loadings < 1
- The relative variance explained by each PC is given by $p_{mi} / \Sigma\, p_{mi}$

Score plot and loading plot - Example

PC1 versus PC2 scores plot



PC1 versus PC2 loadings plot



PCA and noise reduction

- The first PCs explain the most variability of data
- The data noise is concentrated in the last PCs
- Excluding the non-significant PCs can help to “filter out the noise” present in the data

How many PCs should be considered?

Choices:

- Retain a number of components that explain a **determined level** of variance in the data (at least 70-80% of the total)
- Retain a **fixed number** of components
- Kaiser criterion: keep PCs with **eigenvalues >1** (on autoscaled data)

Cluster analysis

- To find groups within a data set, based on the principle for which similar objects are represented by close points in the space of the variables which describe them

- The objects in the same group are more similar to each other than to those in other groups (clusters)

- It can be achieved by various algorithms

- A good clustering method will produce high quality cluster with

  High intra-class similarity.

  Low inter-class similarity.

- The quality of clustering result depends on both the similarity measure used by the method and its implementation.

ENV PRO

Similarity Measures

- different distance measurements can be implemented to evaluate similarity among objects.
- The different measures of distance or similarity are

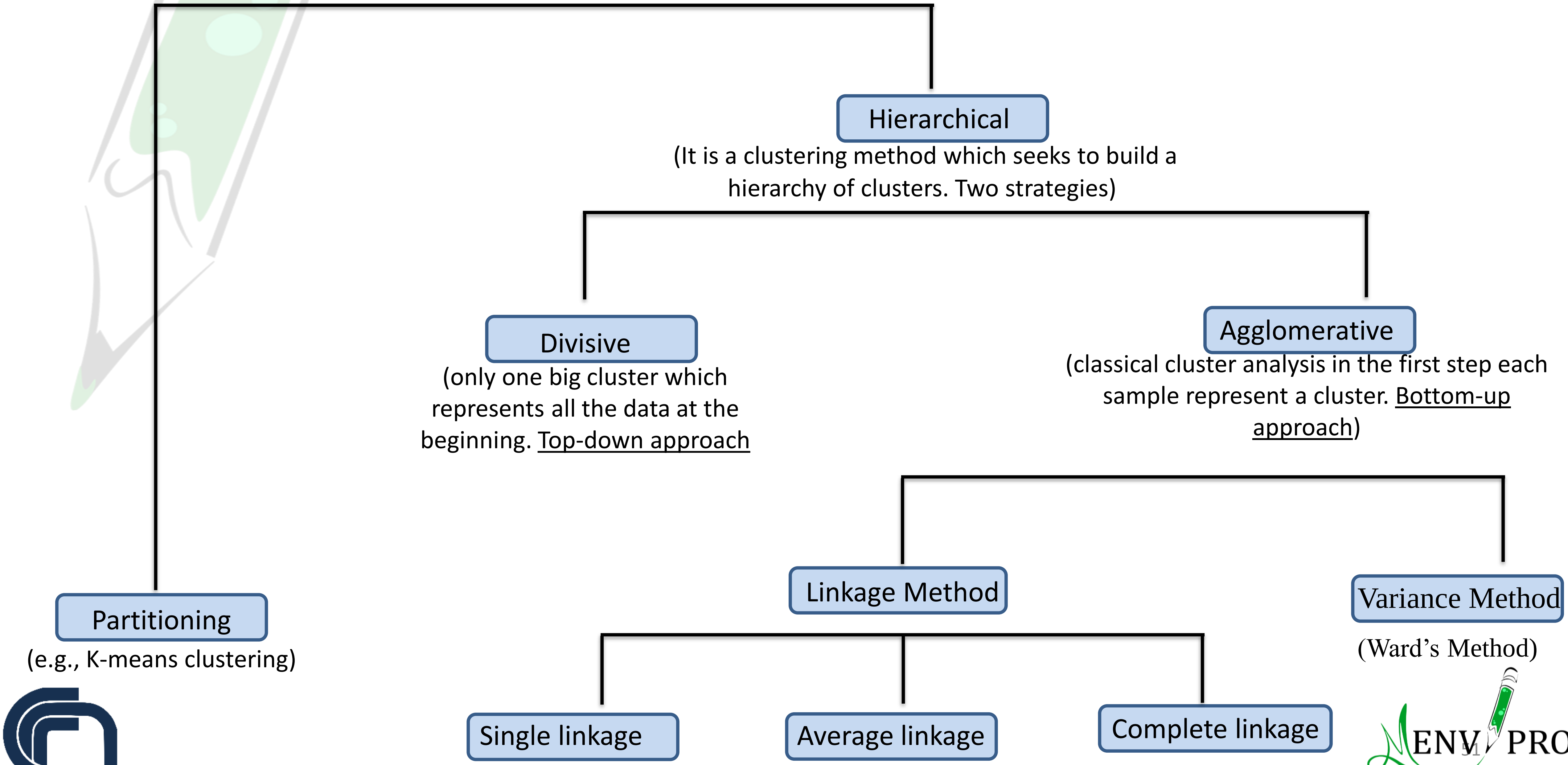
I. Euclidean Distance : $||a-b|| = \sqrt{\sum_i (a_i - b_i)^{\ 2}}$

II. Manhattan Distance: $||a-b|| = \sum_i |a_i - b_i|$

III. Mahalanobis Distance: $\sqrt{(a-b)^T * S^{-1} * (a-b)}$

Where S is the covariance

Types of Clustering Algorithms



Hierarchical

(It is a clustering method which seeks to build a hierarchy of clusters. Two strategies)

Divisive

(only one big cluster which represents all the data at the beginning. Top-down approach

Agglomerative

(classical cluster analysis in the first step each sample represent a cluster. Bottom-up approach)

Linkage Method

Single linkage

Average linkage

Complete linkage

Variance Method

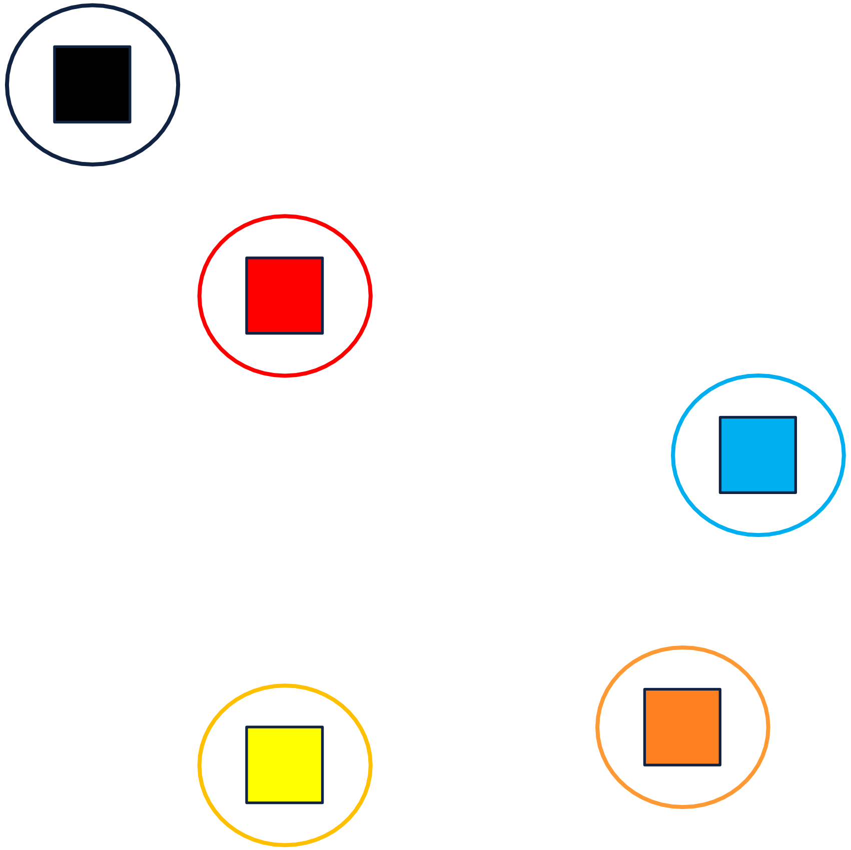
(Ward's Method)

Partitioning

(e.g., K-means clustering)

Hierarchical clustering - agglomerative

1. Assign a cluster to each sample

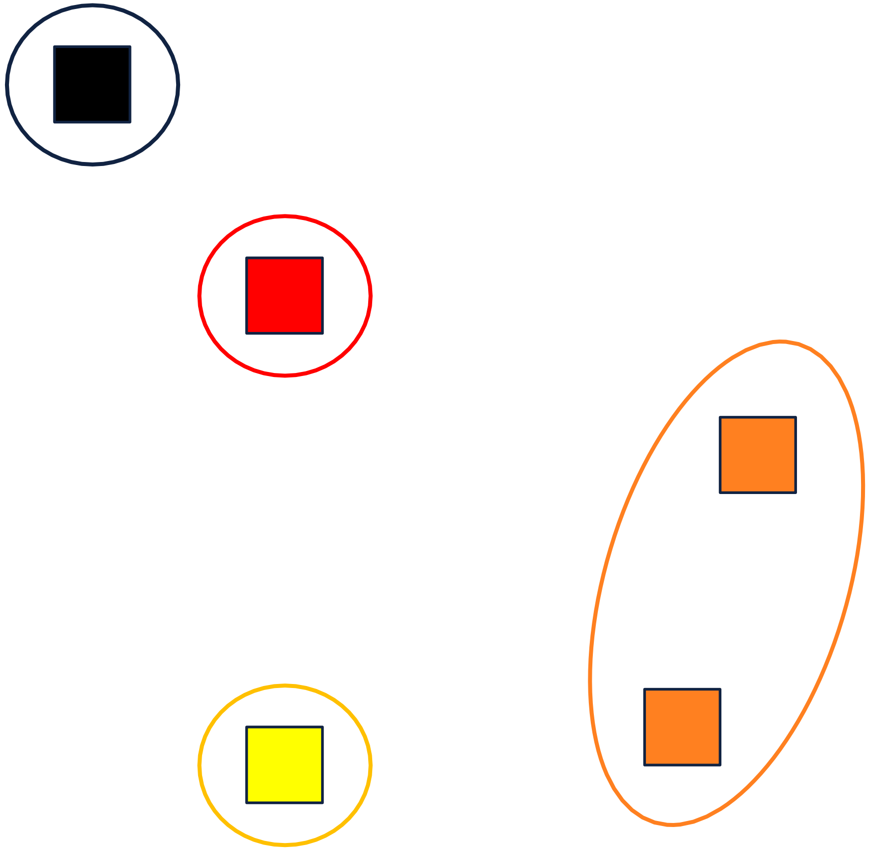


Hierarchical clustering - agglomerative

1. Assign a cluster to each sample

2. Combine the two closest clusters

   (choose your favorite distance method)



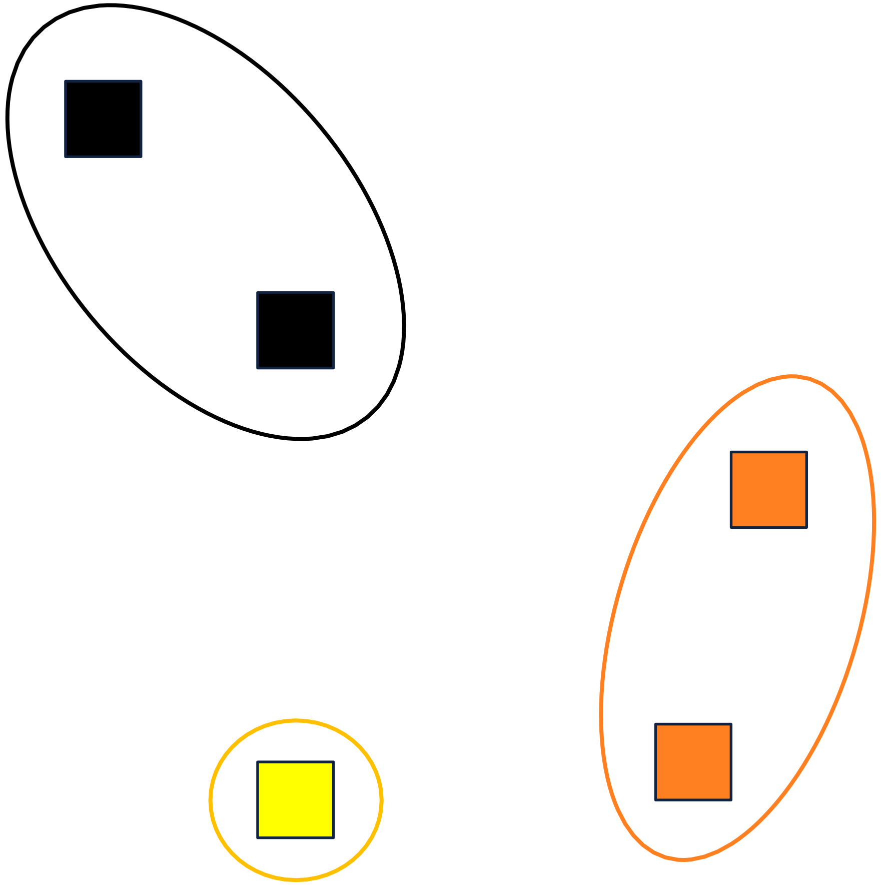
Hierarchical clustering - agglomerative

1. Assign a cluster to each sample

2. Combine the two closest clusters

   (choose your favorite distance method)

3. Repeat step 2.



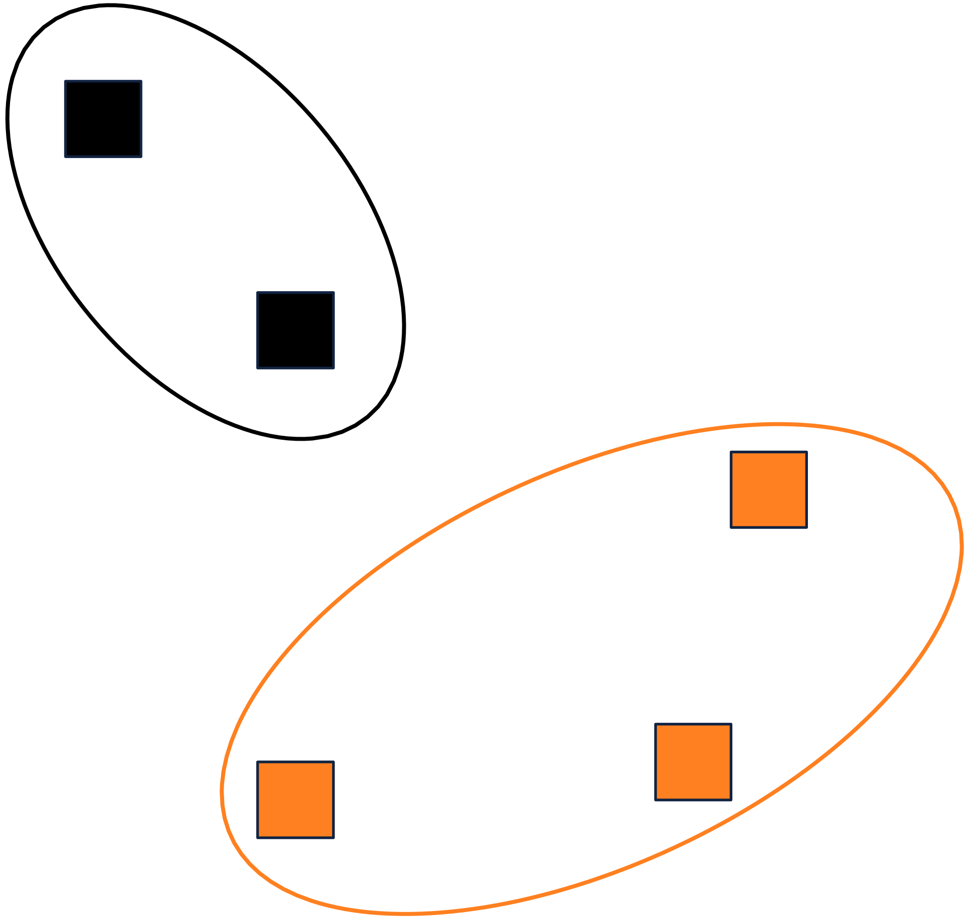
Hierarchical clustering - agglomerative

1. Assign a cluster to each sample

2. Combine the two closest clusters

   (choose your favorite distance method)

3. Repeat step 2.



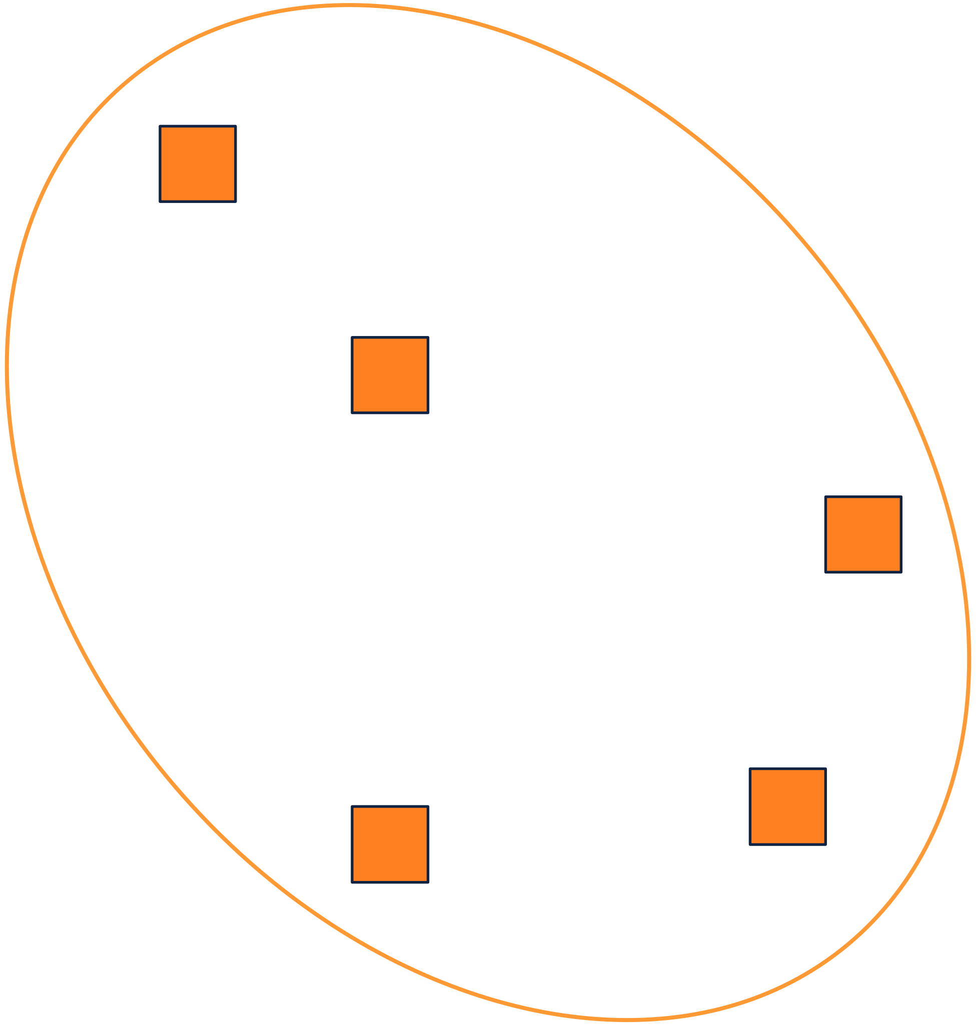
Hierarchical clustering - agglomerative

1. Assign a cluster to each sample

2. Combine the two closest clusters

   (choose your favorite distance method)

3. Repeat step 2.



Hierarchical clustering - agglomerative

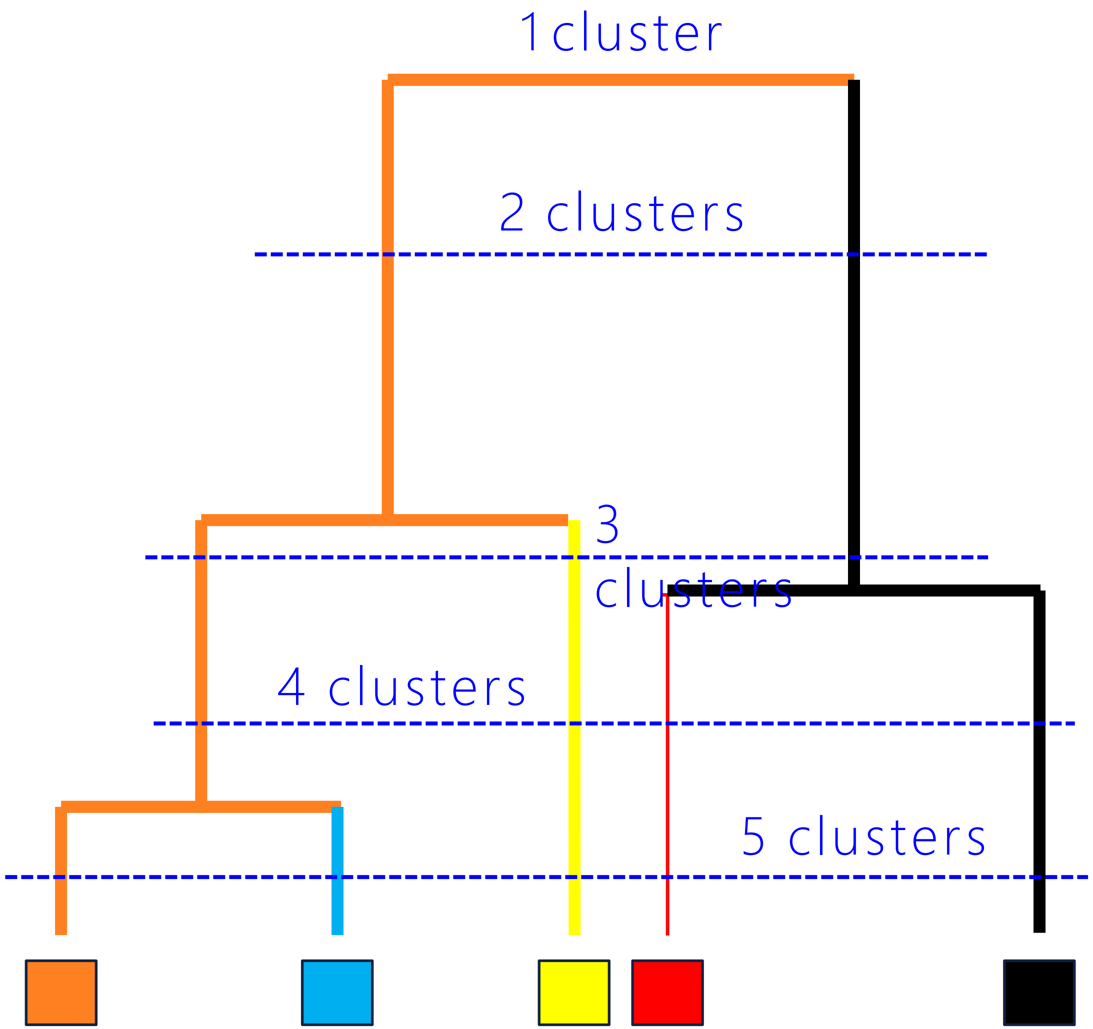
1. Assign a cluster to each sample

2. Combine the two closest clusters

   (choose your favorite distance method)

3. Repeat step 2.

4. Show the results using a dendrogram

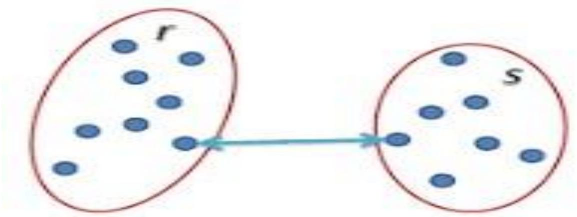


Observations that fuse at the **very bottom** of the tree: **quite similar** to each other Observations that fuse **close to the top** of the tree: **quite different** to each other

How to Define Inter-Cluster Similarity

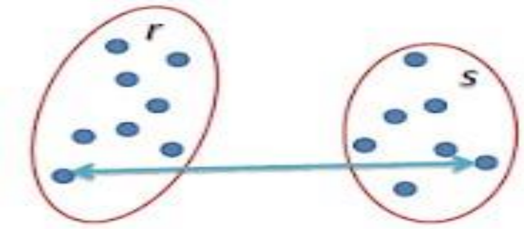
- Single Linkage

The similarity of two clusters is the similarity of their most similar members



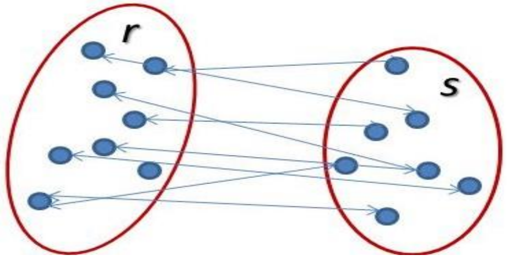
- Complete Linkage

The similarity of two clusters is the similarity of their most dissimilar members



- Average Linkage

The similarity of two clusters is the average distance between each pair of members



- Ward’s Method

The similarity of two clusters is the distance between the sums of squared deviations of the distances of cluster memeber

Proposed statistical and graphic softwares

Collection, organization, analysis, interpretation and presentation of data

Product	Developer	Last version	Open source	Software License	Interface
JMP	SAS Intitute	14.3 - 2019	no	Proprietary	GUI, CLI
Statistica	Dell Software	2015	no	Proprietary	GUI
R	R foundation	3.5.2 - 2018	yes	GNU, GPL	CLI, GUI
Origin	OriginLab	October	no	Proprietary	GUI
Sigmaplot	Systat Software	13.0 - 2014	no	Proprietary	GUI

Co-funded by the
Erasmus+ Programme
of the European Union

Thank you for your kind attention!

Silvia Portarena

silvia.portarena@cnr.it